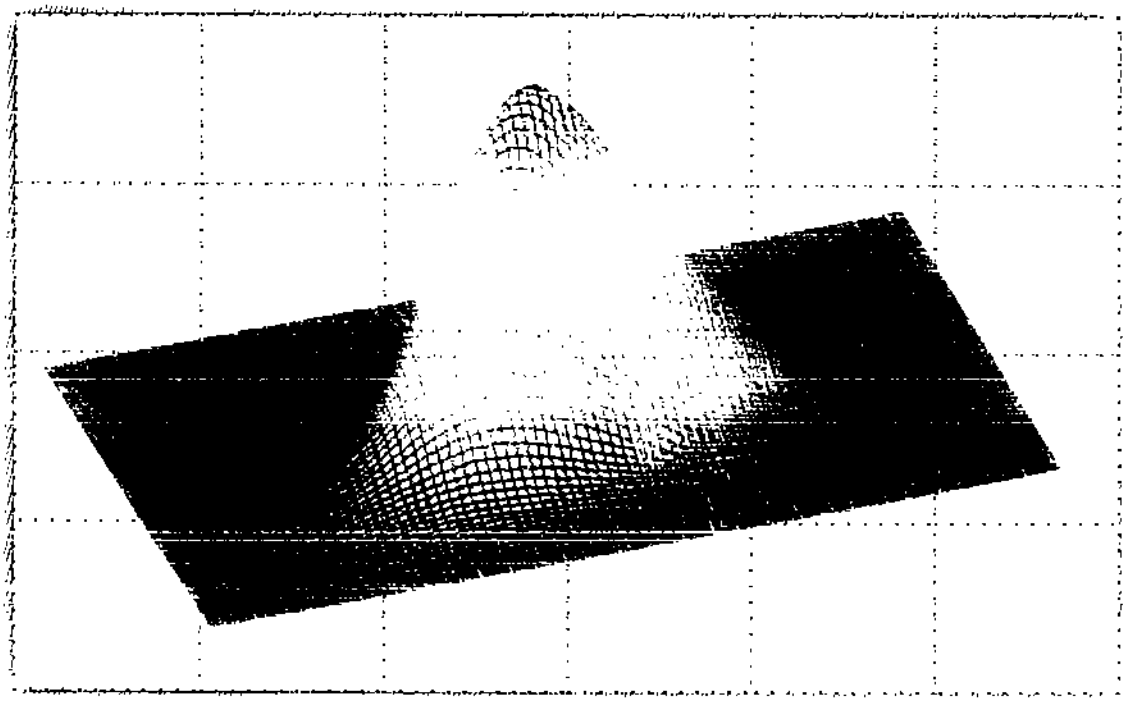
UGSTAT-01
STATISTICAL
METHODS

U P RAJARSHI TANDON
OPEN UNIVERSITY
ALLAHABAD

STATISTICAL *Methods*



Block -I
Data Collection and Its Representation

U.P. Rajarshi Tandon Open
University, Allahabad

UGSTAT-01
STATISTICAL METHODS

Block -I

Data Collection and Its Representation

Introduction

This is the first block on Statistical Methods. It consists of the following three units :

Unit-1 : In this unit, we introduce origin of statistics, its meaning definition and applications along with methods of data collection, measurements and scales.

Unit-2 : In this unit, frequency distribution has been given. Here, Pie Chart, Bar Diagram, Pictograms and leaf chart have also been discussed.

Unit-3 : It deals with graphical representation of data along-with histogram, frequency polygon, frequency curve and Ogives.

Unit -1: Data Collection and Tabulation Structure

Structure

1.1 Introduction

The word "statistics" means status in Latin and "an organized political state" in German and may have been derived from either of them. Initially statistics was known as the science of statecraft and was used by the government to collect the various information needed to administer the state. The great philosopher Chanakya also recognized in his "Arthshaastra" that for an efficient state management, the ruler should keep himself informed about the composition of the sate population with respect to its various aspects such as: literacy, public health, income, and cost of living, etc. In absence of these facts, (which were named statistics later on) the administration may become like groping in darkness.

However with passage of time Statistics did not remain to look simply as the political arithmetic restricted to the study of a state population or to the problems affecting its administration but has assumed quite significant developments by now. In the span started from the later 19th century till date Statistics have taken up unprecedented dimensions and now embrace almost every sphere of nature and human activity. Everyday statistical thinking is becoming more and more indispensable for an efficient citizenship. Through the comparative studies of the qualities and prices of the commodities, even a layman makes use of the statistical methods when as customer one decides as to what quality and from which dealer one should purchase one's daily provisions. There is no newspaper or a periodical these days without having a definite bearing upon statistics. Because of this rapid development and tremendous advancement in recent past, the elementary knowledge of Statistics has become a part of the general education in many advanced and developing countries these days. There is no ground for misgivings regarding practical realization of the dream of **H.H.G. Wells** ***"statistical thinking one day be as necessary for efficient citizenship as the ability to read and write."***

1.2 Objectives

After going through this unit you should be able to

- Know origin of Statstics, its meaning, definitions and applications
- Define universe / Population
- Define statistical problems and limitations
- Know measurements and scales
- Distinguish between discrete and continuous data
- Know methods of data collection, primary and secondary data

1.3 Meaning and Definitions of Statistics

(a) Statistics and Statistical Data :

The word Statistics is used to convey two different senses and is defined differently in each case. One the plural of "statistics" referring to the numerical data collected in an orderly manner with some specific objective in view. A. L. Bowley defines Statistics as numerical statements of facts in any department of enquiry placed in relation to each other. However, Prof. Hrace Secrist defined it as

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

(b) Statistics as Statistical methods or as a tool of analysis:

When the word carries a singular sense it refers to the science of theory and techniques that are used to collect, represent, analyze and draw conclusions from the data. A. L. Bowley defined Statistics as the "science of measurements of social organism, regarded as a whole in all its manifestation". In fact, a number of definitions of statistics denoting singularity are available but perhaps the best one available so far is given by Croxton and Cowden as:

> "Statistics may be defined as the science of collection, presentation,

analysis and interpretation of numerical data."

On the basis of these ideas, we can broadly **summarize that Statistics is a science of**

- Collecting numerical information (data)
- Classification, summarization, organization and analysis of data
- Evaluation of the numerical information (data)
- Drawing conclusions based on evaluation of data

Applications of Statistics

There have been a tremendous growth in the last century that Statistics keeps a role to play in almost every branch of human knowledge; it may be the proper functioning of business and industry, understanding the principles of commerce and economics or the development of the various scientific theories and what not. A few of the multitudes of channels that confront statistics these days are as follows:

- **Statistics in business commerce and industry:** The important areas of business or industry are (a) Production (b) Marketing (c) Personnel (d) Finance and (e) Accounting, where the main functions of Statistics in a practical field of working are planning of operations, establishment of standards and their control. In business problems these statistical functions are conducted either in isolation or in mutual combinations. For example, Statistics is used for quality control in production and is employed for the analysis of sales and marketing in business. Wages and allowances of employees are fixed up on the basis of index numbers. Statistical analysis of costing and accounting data is made for ascertaining profit or loss and for knowing the financial position of the concern at a particular point. Statistical methods are very common and useful to accounts. Audits are done with speed and reliance through sampling. An estimate of the relationship between the cost and volume of production can be made through statistical studies of the past data.
- **Statistical methods** provide a valuable assistance for the study, solution and formulation of **economic policies** on topics like production, distribution of wealth, demand and supply, etc. that no economist can afford to go without their exhaustive studies. **The government intervention in the national economy,** the growth

of large scale entrepreneurial activity and introduction of scientific methods into various parts of business administration has stimulated and contributed to the rapid development of economic-statistics.

- A student of **Physics or chemistry** or of any other **pure science**, while conducting an experiment in the laboratory has necessity to rely upon the application of statistics. An experiment is repeated, its readings vary and in order to reach closest to the accurate result, one has to tabulate them and an average is calculated. In fact **higher studies in every science** need application of the statistical laws like correlation, regression, dispersion, approximation, probability and the tests of significance, etc.
- Innumerable illustrations can be given to show that in **biology** there are frequent applications of statistics. Tests of significance are applied to compare the effects of two or more drugs; the law of probability is employed in irradiation when the cells in the retina of eye are exposed to the light; chart is used to study heart beats through electrocardiograms, and the like. In **agriculture;** the comparison of varieties of seeds or of fertilizers is made through the principles of analysis of variance based on sampling theory. The very fact that **industrial, medical, agricultural, bio statistics** and many more like that are now separate branch of study which speaks of the every expanding scope of statistics and its indispensability in these areas. Statistics also provides a good device of saving time, material and personnel in different studies.

Statistical Applications may broadly be classified under following two disciplines :

(i) Descriptive Statistics :

In descriptive statistics we summarize or describe the data set at hand and evaluate the data sets for patterns and reduce information to a convenient from.

(ii) Inferential Statistics

In Inferential Statistics we use the sample data to make estimates or predictions about a large set of data (also known as population or universe) and test their suitability.

1.4 Universe/Population

The aggregate collection or whole group of individuals or objects possessing certain common characteristics which is of the interest of study is called a population or universe, e.g., population of some college students, population of library books, population of biscuit factories, etc.

Sample is only a part of fraction of population. It is selected with an object of drawing inferences about the various population characteristics (parameters). Each population unit open for sampling is termed as sampling unit and of them units selected in the sample are called sample units.

1.5 Statistical Problems/Limitations

Despite of vast use of Statistics in different dimensions, there are certain limitations of the Statistics and Statistical Methods, which we may call as Statistical problems. These stated as follows –

1. Statistical laws are not exact laws like mathematical or chemical laws. They are derived by taking a majority of cases and are not true for every individual. Thus statistical inferences are uncertain.

2. Statistical methods deal with population or aggregate of individuals rather than individuals. When we say the average height of an Indian is 160 cms, it does not show height of an individual but as found by the study of average or an aggregate of individuals.

3. Statistical techniques apply generally to data which are reducible to quantitative forms. Consequently, the characteristics which can not be

expressed in figures can not be studied satisfactorily. Such characteristics are beauty, goodness, health, intelligence, honesty, etc.

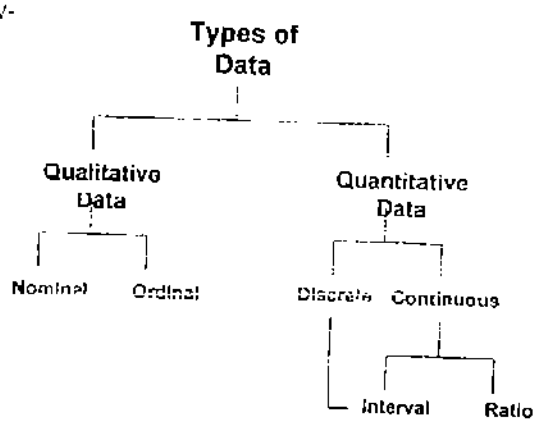
4. Statistical results might lead to fallacious conclusion if they are quoted short of their context. The argument that "in a country 15000 vaccinated persons died of small pox, therefore vaccination is useless" is statistically defective, since we are not told what percentage of the persons who were not vaccinated and had died.

5. Statistical technique is the same for the social and physical sciences, while both are different in nature.

6. Only one, who has as expert knowledge of Statistical methods, can handle the Statistical data property. The data placed in the hands of an inexpert may lead to fallacious results or wrong conclusions.

1.6 Measurements and Scales

Characteristics under study termed as "variable" is a quantitative or numerically expressed qualitative characteristic which varies from one object to another within its domain. Any variable of interest is measured on the units under study to generate the observations known as **statistical data.** Hence we may say that Statistical data refer to the numerical description or measurement of quantitative aspects of things under observations. For example, number of students in a class, number of colleges in a city, temperature, rainfall, etc. Observations or statistical data may be measured according to their classification shown in a diagram below-



1.7 Measurement of Qualitative Data

For a qualitative characteristic called attribute we can simply, observe or note their presence or absence in under observation. There is no natural numeric scale for its measurement. For example, gender, eye colour, beauty, etc. cannot be measured in number or numericals. In such situations we use two types of measurements known as nominal and ordinal scales and resulting into nominal and ordinal data according to their type of measurements.

Nominal Scale

To classify characteristic of people, objects or events into categories under some name is known as nominal scale. For example, gender is classified under name of the male and female, color may be classified as black and white, etc.

Ordinal Scale (Ranking Scale)

Characteristic which can be put under ordered categories measured on ordinal scale are known as ranking scale and data thus generated are known as Ordinal Data. For example, socio-economic status may be measured as low, medium or high status.

Quantitative Data

Statistical data which refers to the numerical description of the character under study of things under observations is known as quantitative data. This description may be in the form of counts or measurements. For example, number of students in a class, and separate counts for various kinds such as male and female students. These counts refer to *discrete types* of variable. The observations may also include measurements as heights and weights, which are referred as continuous variable. The type of variable classifies the type of data as well. There is s natural numeric scale to measure discrete variable or continuous variable such as age, height, weight which is expressible in numbers. It

can be measured according to two scales known as interval and ratio scale of measurement.

Interval Scale

The interval scale is more sophisticated than nominal or ordinal scale. This scale can not only be ordered but the distance between two measurements can be obtained. The distance between these ordered category values are equal because there is some acceptable physical unit of measurement. However, the zero point is arbitrary. It can take continuous or discrete values.

Example :- Fahrenhit (or Celsius) scale for measuring temperature where a temperature of 0^0F does not mean that there is no hear. In fact, there is still some heat at temperatures 10^0F, -20^0F, and so on. Because there is still some heat (the variable being measured) when zero is assigned as the measurement, the zero is not an absolute zero.

Ratio Scale

The highest level of measurement is the ratio scale. This scale has a true zero point and has the "equal ratio" properties. It consists of meaningful ordered characteristics with equal intervals between them. Presence of zero point is not arbitrary and is absolute. It is possible to multiply or divide across a ratio scale. Ratio between two values on the scale is a meaningful measure of the relative magnitude of the two measurements. The reason for the name ratio makes sense to say that a line that is 2.5 cms long is half the length of a 5 cms line. Similarly, it makes sense to say that a 20 seconds is twice the duration of 10 seconds.

1.8 Methods of Data Collection

Once it is decided what type of study is to be made, it becomes necessary to collect information about the concerned study, mostly in the form of data. For this information has to be collected from certain

individuals directly or indirectly. Such a techniques is known as survey method. These are commonly used in social sciences, i.e., the problems relating to sociology, political science, psychology and various economic studies. In surveys the required information is supplied by the individual under study or is based on measurements of certain units.

Types of Data

There are two categories of data namely primary data and secondary data depending upon the method of its collection.

Primary data and its collection

The data which are collected from the units or individual respondents directly for the purpose of certain study or information are known as primary data.

Secondary data and its collection

It is the data which has been collected by certain people or agencies and statistically treated. Now the information contained in it is used again from the records processed and statistically analyzed to extract some information for other purpose. Usually secondary data is obtained from year books, census reports, survey reports, official records or reported experimental findings large scale data can not be collected repeatedly because of the paucity of time, money and personnel. Hence the use of secondary data for certain studies is inevitable. While making use of secondary data, one should always take care of the following points:

(a) One should see whether data are suitable for study.

(b) The source of data should also be viewed, keeping in mind whether at any time, it is reliable or not. If there is any doubt about the reliability of data, it should not be used.

(c) It should be noted that the data are not obsolete.

(d) In case the data are based on a sample, one should see whether the sample is proper representative of the population.

Discrete and continuous Data

Statistical Data may be looked upon as a collection of facts, observation or information in numerical terms on variables under study regarding population/universe or a sample from the universe to achieve the objectives of study or research. That variable which is capable of assuming every possible fractional value within its possible limits (called domain), when measured on different units, is called continuous one; e.g. individual weight, height, age, rod-length, etc. Therefore, continuous data are those which have uninterrupted range of values and can assume either integral or fractional values.

A variable assuming certain specific or the integral values only, when measured, is called the discrete or discontinuous one e.g., the number of members in a family, number of petals in a flower, number of fruits is a baskets and the like. So the discrete data are distinct, separate and invariable whole numbers. Statistical data are also called discrete or continuous data according to the nature of the variable they are associated with.

The statistical methods are applicable only when some data are available. The data can be quantitative as well as qualitative. If the data are qualitative they are quantified by using techniques like ranking, scoring, scaling or coding, etc. The data are collected either by experiment or by survey methods (directly or indirectly) and they are tabulated and analyzed statistically. Whatever may be the resulting value obtained from analysis, proper and correct inferences have to be drawn from these numerical values. These inferences lead to final decision.

Preparation of Tables

Tabulation should not be confused with classification as the two differ in many ways, Mainly the purpose of classification is to divide the data into homogeneous groups or classes whereas the data are presented into rows and columns in tabulation. Hence classification is a preliminary step prior to tabulation. The following steps for the preparation of table are as follows:

1. The shape and size of the table should contain the required number of rows and columns with stub and captions and the whole data should be accommodated within the cells formed corresponding to these rows and columns.

2. If a quantity is zero, it should be entered as zero. leaving blank space or putting dash in place of zero is confusing and undesirable.

3. In case, two or more figures are the same, ditto marks should not be used in a table in the place or original numerals.

4. The unit of measurements should either be given in parenthesis just below the columns captions or parenthesis along with the stubs in the row.

5. If any figure in a table has to be specified for a particular purpose, it should be marked with an asterisk or a dagger. The specification of the marked figure should be explained at the foot of the table with the same mark.

Processing Classification of Data

Before tabulation of primary data, it should be edited for (i) consistency, (ii) accuracy and (iii) homogeneity.

Consistency :

Some information given by the respondent may not be compatible in the sense that information furnished by the individual

either does not justify some other information or is contradictory to earlier one. For example, the total expenditure exceeds the total income reported by the respondent, the number of children mentioned is less than total number of sons and daughters, and then respondent should again be contacted to rectify the mistake so that there may be consistency in data.

Accuracy :

Accuracy is of vital importance. If the data are inaccurate, the conclusions drawn from it have no relevance or reliability. By checking the schedules and questionnaires only a little improvement can be made. For example, if the sum of certain figure is wrong it can be corrected but if the investigation has either made a false report or the respondent has deliberately supplied wrong information about his income, age, assets etc. editing will be of no use. In recent times, checks have been evolved to attain accuracy e.g., by sending supervisors to check work of investigators or reinvestigating a few respondents after a certain gap of time.

Homogeneity :

To maintain the homogeneity, the information sheets are checked to see whether the unit of information or measurements is the same in all the schedules. For instance some people might have reported income per month and some annual income. In such a situation it has to be converted to the same unit during editing. It should also be checked whether or not the same information has been supplied for a particular questions in all the information sheets. The ambiguity arises due to various interpretation of same questions and should be removed.

Once the primary data have undergone the above process it is fit for further analysis.

1.9 Summary

The word "statistics" meaning "status" and "an organized political state" is derived from German or Latin. Statistics as a discipline may be defined as the science of collection, presentation, analysis and interpretation of numerical data and it is applied in various fields like business, commerce, industry, government, biological sciences, social science, agricultural sciences etc. The aggregate collection or whole group of individuals or objects possessing certain common characteristics which is of the interest of study is called a population or universe. The characteristics under study is called variable. Limitation of Statistics is that statistical result are applicable only on group and not on individuals. There are different measurement scales known as nominal, ordinal, interval and ratio scales. Depending upon the finite or infinite number of values taken by variables they are classified as discrete or continuous. Data collected directly by the method of interview, measurements or questionnaire, etc is called primary data whereas those taken from previous records are called secondary data.

1.10 Self Assessment Questions

1. Give different senses in which word "statistics" is used.
2. Describe the scope of statistical methods and specify their limitations.
3. Which of the following are statistical statements? Give reason.
   - Shakespeare was a great poet.
   - The average age of students of this Institute is 20 years.
   - The production of sugar in a certain district was quintals per acre in a particular year.
4. Comment on the following
   - Statistics can prove anything.

- Figures won't lie but lairs figure.

5. What is a statistical data and how it is classified ?

6. Differentiate between the following :
   - Quantitative and qualitative data
   - Discrete and continuous data
   - Nominal and ordinal scale
   - Interval and ratio scale
   - Primary and secondary data

7. Discuss various methods of measurements of data

8. Write a note on methods of collection of data?

9. State briefly the advantages of "secondary data" over "primary data".

10. What are the precautions to be taken while handling the secondary data?

1.11. Further Readings

1. Goon. A.M. Gupta, M. K. and B. Dasgupta : Fundamental of Statistics, vol. One, the World Press Pvt. Ltd., Calcutta.

2. Yule. G.U., and Kandall, M.G.: An Introduction to the theory of Statistics. Charles Griffin & Company Limited.

3. Weatherburn, C.E. : Mathematical Statistics.

Unit-2: Representation of Data-1 (Diagrammatical representation)

Structure

2.1 Introduction

We have seen in the last unit how Statistics and Statistical Methods provide a valuable assistance for the study, solution and formulation of different kind of problems in almost all spheres of human activities. The statistical data are collected either by experiment or by survey methods (directly or indirectly). The way data is coming as and when observed, these data follow no order and are offered perhaps the way originally reported. There may be a complete lack of any systematic arrangement by size or sequence. Such unorganized data are known as *raw data or ungrouped data.* The data collected through surveys or experiments is the raw data and will be in a haphazard and unsystematic

form. Such a data is not appropriately formed to draw right conclusions about the group or population under study. Hence it becomes necessary to arrange or organize data in a form, which is suitable for identifying the number of units belonging to a more classified group for comparison and for further statistical treatment or analysis of data.

Statistical data which usually refers to the numerical description of the character under study of units or things under observations is known as quantitative data and may be in the form of counts or measurements. For example, number of members in a family, and separate counts for various kinds such as male and female. These counts refer to *discrete type* of data. The observations may also include measurements as heights and weights, which are referred as *continuous data*. There is a natural numeric scale to measure discrete variable or continuous variable such as age, height, weight which is expressible in numbers.

The placement of these data in different homogeneous groups, formed on the basis of some characteristics or criterion is called classification or tabulation of data leading to better understanding and statistical analysis.

Let us consider an example where the raw data is collected about the group or population under study. For instance the people may be divided into different age groups like <10, 10-20, 20-30, 30-40, etc. Or may be classified according to their monthly income (in Rupees) like <500, 500-1000, 10012000, etc. Further these classified data can be presented in the form of well arranged tables.

These tables depict clearly the values or number of units possessing the required characters or belonging to specified classes.

Time series data

Another type of classification is the time series data in which data or the derived value from data for each time period is arranged

chronologically.

2.2 Objectives:

After going through this **unit** you should be able to know about –

- Frequency distribution
- Diagrammatical representation
- Pie chart
- Bar diagram
- Divided bar diagram
- Percentage bar diagram
- Pictogram
- Leaf chart

2.3 Frequency Distribution

The premiere of data in form of frequency distribution describes the basic pattern which the data assumes in the mass. Frequency distribution gives a better picture of patterns of data if the number of items is large enough.

From a frequency array, it is not possible to compare characteristics of different groups. Hence for this, the classes are established to make the series of data more compact and understandable. The width of a class i.e. the difference between the upper and the lower limit of the class is termed as class interval. Once the classes are formed, the frequencies for these classes from raw data are expedited with the help of little slanting vertical strokes called tally marks. A bunch of four tally marks is crossed by the fifth to make the counting simpler. The whole process is as follows-

Defining the task

Let us assume that we have a set of data collected through a sample survey, which consists of a given number of observations on a certain quantitative variable. They differ in magnitude but in the manner presented these raw observations do not exhibit any sensible pattern and therefore the first and foremost task in dealing with such data is to arrange them in right format. The format should aim at

(i) Organizing data in a manner that these become easy to read, understand and assimilate, and

(ii) Summarizing data in a way that the basic trends and broad variations come to the fore and get highlighted.

When presented in the resultant form, a researcher or an analyst gets a better grasp of the data. It facilitates a more efficient data analysis, which helps quicken the process of decision making.

Example

Consider the following raw data given in table which refers to weekly earnings of 80 female workers engaged in weaving trade at Surat during a particular year.

Weekly earnings of 80 female workers engaged in weaving trade at Surat

1052	1088	1077	1078	1089	1089	1082	1084	1088	1090
1099	1101	1102	1055	1063	1073	1078	1113	1086	1089
1080	1095	1092	1103	1118	1098	1097	1081	1061	1080
1083	1079	1111	1064	1056	1068	1055	1073	1075	1083
1085	1086	1083	1090	1105	1090	1069	1058	1072	1073
1086	1071	1070	1065	1059	1080	1084	1085	1075	1064
1087	1091	1108	1094	1097	1093	1107	1094	1082	1116
1085	1070	1076	1069	1061	1114	1089	1074	1105	1082

Preparation of Frequency Distribution (Tally Method)

To make a frequency distribution table for above example by using tally marks we proceed as follows-

1. Obtain the range of the distribution as the difference between the lowest and the highest observation(s). For the data listed in table 1052 is the lowest and 1118 is the highest observation, with (1118- 1052=) 66 as the range of the distribution.

2. The range is then divided into an appropriate number C, which represents the width of the class interval. This also determines the number of class interval k among which individual observations are distributed. If C= 10, the range 66 is divided by 10 resulting in 6.6 n rounding to the next higher digit, it gives k=7 class intervals.

3. After completing step 2), all individual observations in the original data are picked up one by one and a tally bar is marked opposite the class in which a particular observation falls. For example, in table an observation 1052 lies in the class (1052-1059) so that a tally bar is marked against this class. This has to go on till all the observations have been recorded by making tally marks.

4. Finally, tallies marked against each class are counted and their total number recorded under a separate column heading frequency, as is column *(3)*. For convenience in counting the number of tallies entered in each class, every fifth tally mark crosses the earlier our tallies diagonally from top to the bottom. Adding all class frequency yield a number 80 equal to the total number of observation (N) so recorded.

5. Frequencies obtained in column 3 may be expressed as present class frequencies as shown in column 4.

Frequency distribution of weekly earnings of 80 female workers engaged in weaving trade at Surat (Tally Method)

Class – limits 'C' (1)	Tally Marks (2)	Frequencies 'f' (3)	Percentage class frequencies (4)
1050 – 1059	𝍸 I	6	7.50
1060 1069	𝍸 𝍸 I	9	11.25
1070 1079	𝍸 𝍸 𝍸 I	15	18.75
1080 – 1089	𝍸 𝍸 𝍸 𝍸 I	25	31.25
1090 1099	𝍸 𝍸 III	13	16.25
1100 – 1109	𝍸 II	7	8.75
1110 - 1119	𝍸	5	6.25
Total		80	100

The distribution constituted by column 1 and 3 in the above table is known as frequency distribution. It gives the number of women according to their weekly earnings .For example 6 women's earning is between Rs1 050 to Rs. 1059; 9 women's earning is between Rs. 1060 to Rs. 1069 and so on. This frequency distribution has helped to understand and analyze the haphazard data in a systematic manner which is easy to handle for further treatment.

Smoothening of a grouped Distribution (Inclusive and Exclusive type Class Intervals)

When the upper limit of the previous class is not as the lower limit of the following class. as in the above example. and both the class limits are included in the same class are called inclusive type class intervals. In such case the classes do not constitute the continuous distribution and has to be made continuous. The simplest way to do this

is to find the difference of the upper limit of the preceding class and lower limit of the succeeding class. Subtract half of the difference from the lower limit of each class and add the same from its upper limit. Continue this process for all the classes. In the resulting class intervals the upper limit of the previous class is same as the lower limit of the following class, and only lower class limit is included in the corresponding class not both the limits as in case of inclusive type of class intervals and are therefore called exclusive type class intervals.

Frequency distribution (Inclusive type class intervals) of the Weekly earnings of women

Weekly earning (class limit)	Number of women (Frequency)
1050 -- 1059	6
1060 – 1069	9
1070 – 1079	15
1080 1089	25
1090 – 1099	13
1100 - 1109	7
1110 - 1119	5
Total	80

The given distribution is not continuous as the upper limit of the preceding class is not the same as lower limit of following class. Hence it is smoothened. The difference between 59 and 60 is 1. Therefore, 0.5 is to be subtracted from the lower limit of the classes and 0.5 is to be added to the upper limit of the classes. Since the difference is constant the same quantity is subtracted and added in all the classes.

Frequency distribution (Exclusive type class intervals)

of the Weekly earnings of women

Weekly earning . (class limit)	Number of women (Frequency)
1049.5 – 1059.5	6
1059.5 – 1069.5	9
1069.5 – 1079.5	15
1079.5 - 1089.5	25
1089.5 – 1099.5	13
1199.5 - 1109.5	7
1109.5 – 1119.5	5
Total	80

Open End Classes

An open end class is a class taking one limit. Generally it is the lowest class taking the lower limit and highest class taking the upper limit. For instance, in an age group distribution, the lowest class is taken as less than five (<5) and highest-class as more than seventy (>70). Open end classes make it possible to accommodate values which are at large gaps without increasing the number of consecutive classes. However, open end classes should be avoided as far as possible. Open ends create pt6blem in processes like computation and graphical representations.

Cumulative Frequency

The frequencies may be added up or cumulated on either from top to bottom (on the less than basis) or from bottom to the top (on more than basis). Cumulative frequencies less than type are obtained by adding successive frequencies from top to bottom as given in col. (4).Those of 'more than type' cumulative frequencies are obtained by adding successive frequencies from bottom to top as given in col. (5) as shown in following table. The less than type cumulative frequencies correspond to the upper limit of the class whereas more than type cumulative frequencies correspond to the lower limit of the class.

Cumulative frequencies

Weekly earnings (class limit)	Number of women (Frequencies)	Cumulative frequencies (less than type)	Cumulative frequencies (more than type)
1049.5-1059.5	6	6	80
1059.5-1069.5	9	15	74
1069.5-1079.5	15	30	65
1079.5-1089.5	25	55	50
1089.5-1099.5	13	68	25
1199.5-1109.5	7	75	12
1109.5-1119.5	5	80	5
Total	80		

In the above example cumulative frequencies help in finding out total number of women whose earning is less than Rs. 1059.5 is given by 6 whereas total number of women whose earning is more than Rs. 1049.5 is given by 80. similarly total number of women whose earning

is less than Rs. 1069.5 is given by 15 whereas total number of women whose earning is more than Rs. 1059.5 is given by 74 and so on.

2.4 Diagrammatical Representation

How so ever informative and well designed a table may be, the pictorial representation of data is definitely a better tool for conveying details of the data to the common man in a simpler and more understandable manner. The figures given in tabular form as such are not easily intelligible because of their dull, confusing and dispelling nature. If they are large in number or in size, then their study needs much time and brings a strain upon the mind. Diagrams on the other hand, are attractive and catch the attention of the reader by explaining and exposing the significant facts given in the data in a visual and summarized form. They have a more lasting effect on the brain. When data of two items are compared with one another it is always easier through diagrams and graphs. It is for this reason that the government, various business houses and institutions are producing popular versions of their important statistics these days in the form of multi-colored booklets full of pictures, geometrical figures, curves and maps, etc. So it is very useful to represent statistical data by means of a diagrams and graphs which make of the unwieldy data intelligible and convey to the eye the important characteristics, general tendency and trend of the observations. It is now an essential part of the analysis and presentation process of statistical data. The four basic purposes of diagrams/graphs are to

(1) Compare "proportions & relative changes" in data

(2) Show trends/ tendencies of data

(3) Study how response changes over time in given set of

data

(4) Indicate how one variable relates with one another.

For a graph, a proper title, labeling for the axes and units of measurement are important. A good graph is designed so that it gives brief description at a glance. Similar descriptions should be provided for diagrams also.

However, there are certain disadvantages also. Graphs do not give accurate measurements of the variables as are given by tables. The numerical value can be obtained to any number of decimal places but from graph it can not be found, say to 2^{nd} or 3^{rd} places of decimals. Another disadvantage is that it is very difficult to have a proper selection of scale. By different scales it is possible that the facts may be misrepresented.

Different Categories of Diagrams

In this section we are going to deal with the following diagrams and graph to represent different types of statistical data

- Bar diagram
- Divided bar diagram
- Percentage bar diagram
- Pie Chart
- Pictogram
- Leaf chart

2.5 Bar Diagram

As the name indicates the bar diagrams are the simplest one.. dimensional diagram. A bar diagram is a visual display used to compare

the frequency of occurrence of different characteristics of categorical data. They are particularly used if items of the different classes are not component parts' of the whole but are related with each other by their possession of some common characteristics. Bar height commonly represents a count of cases or frequencies for each category, a percentage of the total number of cases, or a function of another variable (e.g.. the mean value for each category). Here thickness of the bar has to do nothing with the interpretation of the figures for which only the length or height is taken into account.

Bars are spaced at an equal interval. As far as possible they should be placed in ascending or in descending order of their lengths, allowing half bar width between bars and at each end vertical scale should start only with "0" only.

The scale should be mentioned on the diagram but it should be quite a convenient scale.

The main purpose of the bar graph is to

1. Compare groups of data
2. Make generalization about the data quickly

Example

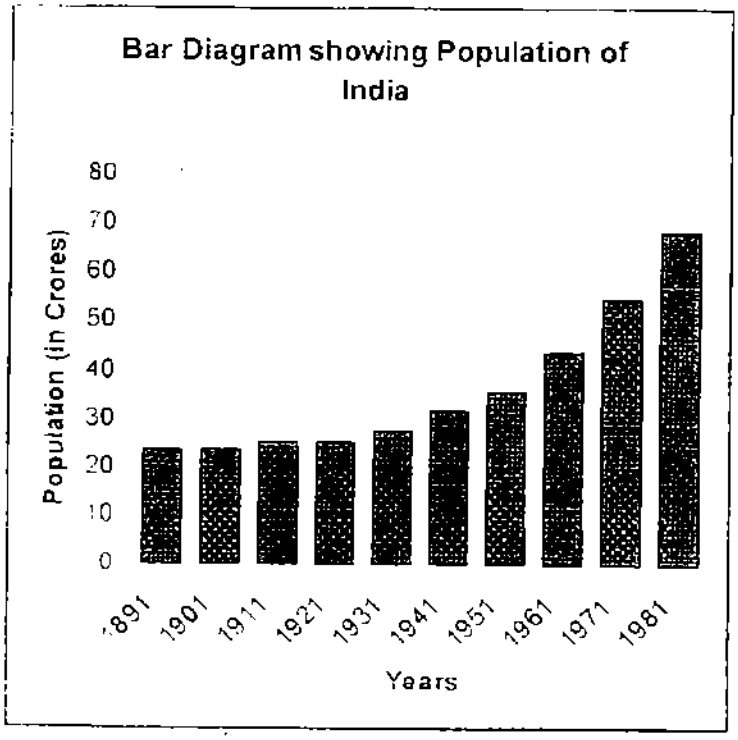
The bar diagram is shown in figure for the following data regarding population of India given in the table.

Table

Year of Census	Population(in crores)
1891	23.59
1901	23.83
1911	25.20
1921	25.13

1931	27.89
1941	31.86
1951	36.10
1961	43.91
1971	54.82
1981	68.38

showing Bar data given in



2.6 Multiple Bar Diagram

They are extended form of the simple bar diagram. Here more than one aspect of the data is presented simultaneously. Each aspect is shown with different shades or colours. The bars of one group are separated from other groups by putting them adjacent. Also the multiple

bars can be shown as placed on one another without loss of clarity to save space.

These diagrams are very useful for comparison between two or more phenomena by representing them with different bars having different shades or colors. An index explaining shades /colors and scales used should be shown in the diagram. Bars may be horizontal or vertical .The space between bars representing the components of the same total is taken smaller than space between bars of the different sets of total. However, the totals in themselves are not easily comparable here as though the simple bar diagrams. These diagrams are meant for comparing two or more sets of interrelated data over different points of time, place or categories, etc.

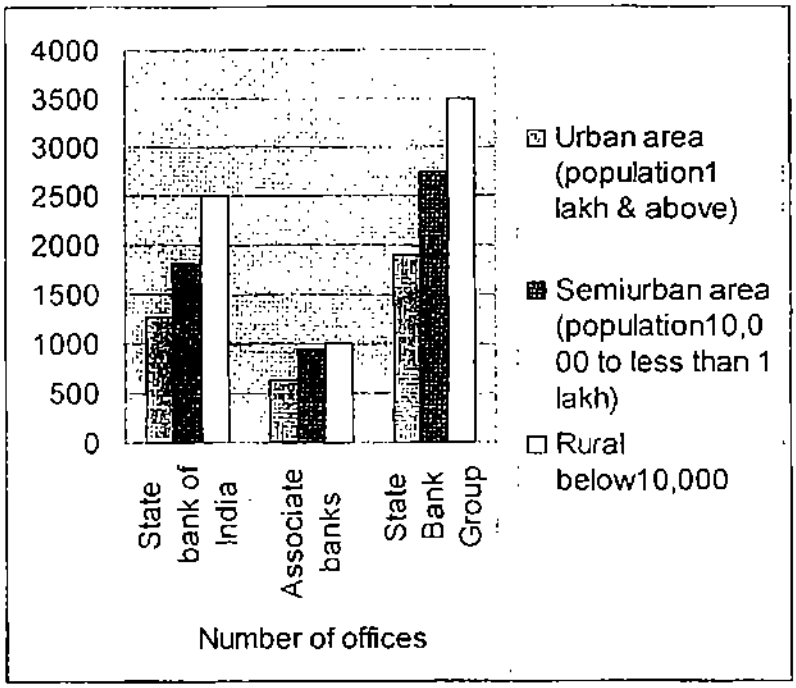
Example

Following data shows State Bank Group population-wise distribution of Indias offices at end of 1980).Prepare a multiple Bar Diagram.

Table - Population wise distribution of Indian Offices of SBI

Population Category	Number of offices		
	State bank of India	Associate banks	State Bank Group
Urban area (population 1lakh & above)	1270	632	1902
Semi urban area (population 1 0,000 to less than 1 lakh)	1807	940	2747
Rural below 10.000	2492	999	3491

Figure showing Multiple Bar Diagram for above table



2.7 Subdivided Bar Diagram

When it is desired to show the aggregates and their division into various components, the bars are drawn proportional in length to the totals and are subdivided into ratios of their components. Each subdivided part of the bar will correspond in size to the value of the item, it represents. Such diagrams are called Subdivided Bar Diagram.

While preparing these diagrams, it must be observed that the arrangements of the various components remain identical for all the bars to avoid confusion and keep the diagram readily distinguishable. As usual, different shades or colors are to be used for representing different components of the total but shades of each component will remain the

same for all the bars. Index of shades and scales used should be shown with the diagram.

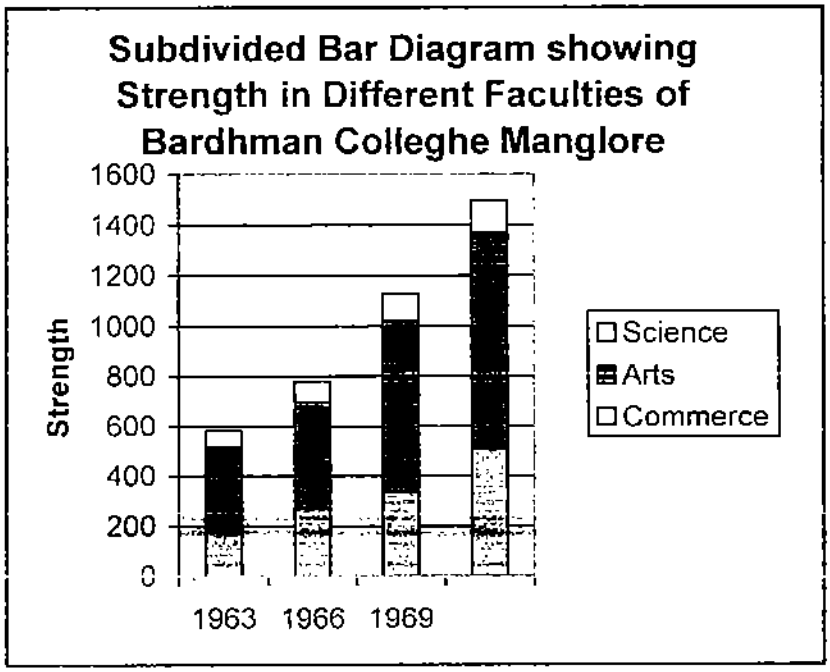
Example

Draw a subdivided bar diagram for the following data given in table.

Table

	Strength of students of Bardhman College Manglor in the year									
Faculties	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972
Science	177	185	260	270	296	341	338	398	459	507
Arts	340	414	417	426	470	681	684	696	779	862
Commerce	65	85	82	84	84	109	106	87	103	125
Total	582	684	795	780	850	1131	1128	1181	1361	1494

Figure - showing Sub-divided Bar Diagram for table



Thus We see from figure that these subdivided bar diagrams denote not only the variation in the total of the values of given characteristic but changes in the component parts of the total are also exhibited.

Limitations

Since the components of the bars do not start from the same scale value, so they do not remain easily comparable in their size across-sections. Here individual bars are to be studied separately and properly for the inter component comparisons.

2.8 Percentage Subdivided Bar Diagram / Percentage Bar Diagram

If the purpose of the graph is only to show the proportionate composition of the totals with respect to their component parts, it is best served by a Percentage Subdivided Bar Diagram or simply by a Percentage Bar Diagram. Here all the totals are equated to 100% and are represented by the bars of the same length. The component figures are expressed as percentage of the totals to obtain the necessary length for them in the full length of bars representing totals. The other rules regarding the shades, index and thicknesses are the same as those of simple and multiple bar diagrams. Here the comparison within the original observations is not possible. But, it remains easier to plot cumulative percentages from the bottom.

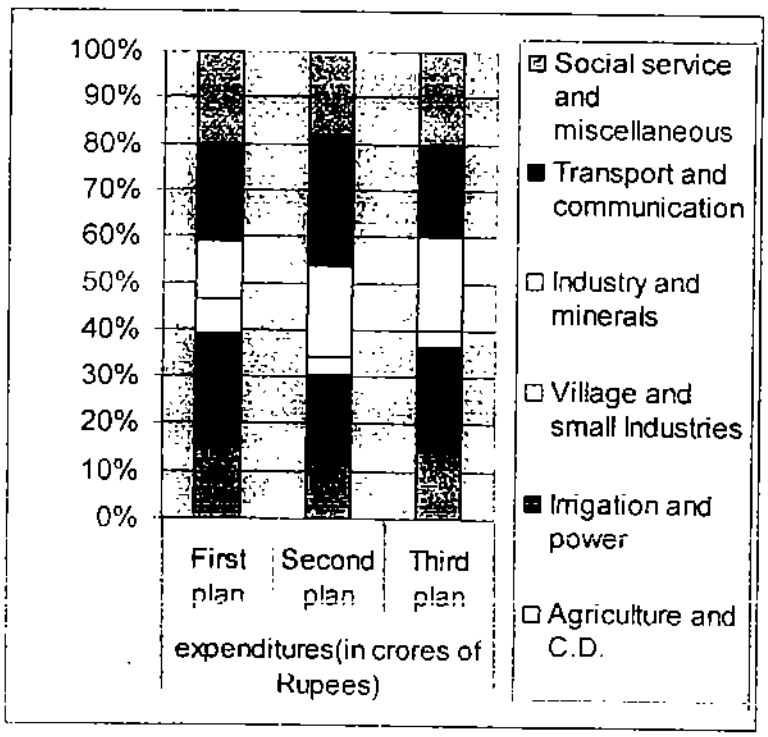
Example

Following are the expenditures (in crores of Rupees) on various heads incurred in the 1st, 2nd, and 3rd five year plans given in the following table. Draw a percentage bar diagram.

Table

Subject	Expenditures (in crores of Rupees)		
	First plan	Second plan	Third plan
Agriculture and C.D.	361	529	1068
Irrigation and power	561	865	1662
Village and small Industries	173	176	264
Industry and minerals	292	900	1520
Transport and communication	497	1300	1486
Social service and miscellaneous	477	830	1500
Total	2361	4600	7500

Figure showing percentage sub divided bar diagram for table



2.9 Pie Chart

Circles are preferred than rectangles bars when the difference between totals to be compared is larger. Circles are drawn with their radii in proportional to the square roots of the values they represent. All the centers of the circles must lie in a straight line. If totals are the sum of the various components. then each circle may be divided into as many segments as are the components in its corresponding total. The area of the segment has the same percentage to the total area of the circle as the represented valve has with its total figure. We know that the sum of the angles round the centre of a circle is 360 degrees.

Pie graph displays percentages. The circle of a pie graph represents 100%. Each portion that takes up space within the circle stands for a part of that 100%.

$$\text{The angle of the sector} = \frac{\textit{Value of the represented part}}{\textit{The whole quantity}} \times 360^{0}$$

Pie-diagram is also known as Circular diagram due to its shape.

Example

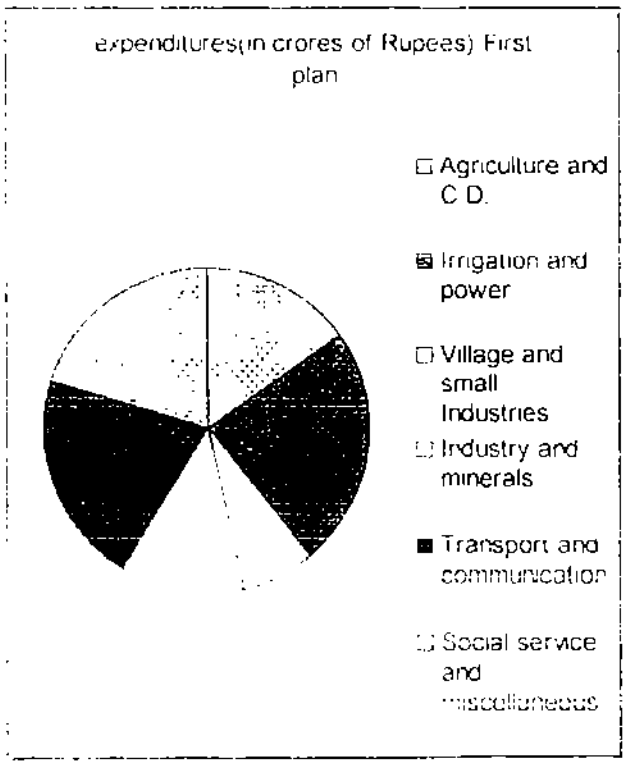
Represent them by by a pie diagram the expenditures (in crores of Rupees) on various incurred in the first five year plan given in table.

Since total expenditures is Rs. 2361 Crores so the angular measures of the individual components are obtained as follows-

Table

Subject	expenditure (in crores for Rupees)	
	First plan	degrees
Agriculture and C.D.	361	55
Irrigation and power	561	86
Village and small Industries	173	26
Industry and minerals	292	45
Transport and communication	497	76
Social service and miscellaneous	477	72
Total	2361	360

Figure showing Pie Chart/diagram for adjacent table



2.10 Pictograms

The device of pictures is being profusely used now for comparing statistical data. The pictorial representation of facts are very often used in various exhibitions, propaganda posters, etc. They present dull masses of figures in an interesting and attractive manner through the objects of daily picture. The image of the entire data is fixed in the mind of the observer by a mere glance at the picture. Relationship between figures and their comparison can be studied through pictograms much more easily than by studying huge mass of numerical data.

2.11 Leaf Chart

Drawing a histogram, discussed in unit III, can often be quite tedious and an alternative method can be employed when the raw data (original non-grouped) are available. Table shows the ages at which 21 females were admitted to hospital with a hip fracture.

Table showing the ages of females admitted with a hip fracture.

53	76	84
62	78	86
71	82	87
71	78	85
67	84	87
73	84	94
73	84	98

A histogram can be drawn for these data, but the same end result can be achieved with the stem-and -leaf diagram. If the tense part of the age were taken as a "stem" on which the unit parts of the age were to be attached like a leaf, would be obtained.

The stem is written down in the first column and usually a vertical line is drawn to separate the stem from the leaves. Here the leave is the unit's part of the age and the leaves belonging to each stem are put in the next columns. Thus there is only one patient in her 50s at the age 53 and the digit "3" goes in the first leaf column on to the stem "5".there are two age (62 and 67) that belong to the second stem of "6", and the units '2' and '7' go in the first and second leaf columns. In this way, the leaves are added to the stems to give the complete diagram.

Stem-and -leaf diagram of the ages (years) of 21 females patients with a hip fracture.

5	3								
6	2	7							
7	1	3	3	3	6	8	8		
8	2	4	4	4	4	5	6	7	7
9	4	8							

If the diagram is turned so that the stem is at the bottom it can be seen that a crude histogram of data has been created. The class corresponds to the range of values implied by the stem -in this example. from 50 to 59 years. 60 to 69 years. etc. As long as the digits of the leaves are written in equal width columns, the number of observation in each class is given by the length of the row, which corresponds to the height of the bar of the histogram.

The stem and leaf diagram is quite simple to draw and the digits for the leaves need not be put in ascending order. In fig. for instance, the '8' (for 98) could have been put in the first leaf column of the '9' stem

and the '4' (for 94) could have been put in the second column. Thus once the appropriate stem~ has been chosen, the leaves can be filled in easily by just going through the data without even having to order them.

For a given set of data, choosing an appropriate stem is some times a process of trial and error, but there are a few tricks that can ease the task. Usually the number of stems is between five and twenty but this depends on having a sufficient number of values for each stem. Note that it is not

possible to create three stems within a single tens digit, since the number of possible different leaves on each stem must be the same.

2.12 Summary

For a better presentation and efficient analysis statistical data are classified, summarized and tabulated in the form of Frequency distribution using tally marks. Data may be grouped in inclusive or exclusive type of class intervals. This frequency distribution can further be treated and cumulative frequencies may be obtained. For better understanding these frequency distributions can be shown on graphs as histogram, frequency polygon, frequency curve and ogives . Statistical data are diagrammatically represented as chart, bar diagram, divided bar diagram, percentage bar diagram, pictogram, and leaf chart for facilitating analysis and comparisons of data over person, place and time giving lasting and eye catching effects.

2.13 Self Assessment Exercises

1. Mention the methods generally used in the collection of statistical data with precautions to be taken.
2. What are different parts of a table? How a frequency distribution table is prepared?

3. What do you mean by classification and tabulation of data?

4. What do you understand by diagrammatic representation of data? What are its main advantages and disadvantages?.....

5. Discuss various methods of graphical representations stating the situations where they may be optimally used. Also state their limitations precautions to be taken.

6. For the data given below, draw a) a simple bar chart for GDI as percent

to GDP, and b) a single chart showing the share of public and private GDI as percent to GDP.

GDI at Current Market Prices (as percent to GDP)

	1983-84	1984-85	1985-86	1986-87	1987-88	1988-89
GDI	23.1	26.1	27.2	24.6	26.2	23.4
Public	8.2	8.8	7.6	7.0	6.7	6.6
Private	13.0	14.8	18.9	14.9	14.7	15.2

7. Using the data given below, draw a) component chart for absolute figures, b) component chart after obtaining percent data, and c) pie charts for percent data for 1975-76 and 1980-81.

Gross Savings at Current Prices (Rs Crore)

Year	Household	Pvt. Sector	Public Sector	Total
1975-76	28058	5243	7642	40943
1976-77	41567	5125	7539	54231
1977-78	65519	5789	6981	78289
1978-79	78248	7983	7803	94034
1979-80	82145	11580	7562	101287
1980-81	100453	15642	5423	121518

2.14 Further Readings

(1) Goon, A.M., Gupta, M.K., Dasgupta, B. : Fundamentals of Statistics, Vol. One, The World Press Pvt. Ltd., Calcutta

(2) Yule, G.U. and Kandall, M.G. : An Introduction to the theory of Statistics. Charles Griffin & Company Limited

(3) Weatherburn, C.E. : Mahematical Statistics.

Unit-3 Representation of Data – 2 (Graphical Representation)

Structure

3.1 Introduction

In general, the statistical data are unwieldy and as such its various features cannot be understood clearly and readily, at a glance. It has to be reduced in a suitable form. The representation of the data in a tabular form by a frequency distribution is one of the techniques to achieve this objective. Normally the frequency distribution is not able to highlight various salient features of the data. A graphical representation of the frequency distribution is a powerful tool of data representation and interpretation. Hence, the data is represented by way of lines, curves, dots and bars etc. on a graph paper with variable values being put on the X-axis and frequencies on the Y-axis. The graphical representation of the data is an attractive and impressive way of representation that has a more lasting effects on the mind of the human beings than the tabular

form representation of the data. The shape of the graph provides easy answers and ideas regarding the variation of data, Skewness, peakedness at the top of the frequency curve, modes, extremes, outliers, spread of the data etc. inherent in the distribution of the data. Accordingly, frequency distribution graphs serve as an effective tool of a quick analysis and effective comparison between two or more distributions. The pattern of variations and the points of contrast become quite obvious when the graph of one distribution is superimposed over the other.

3.2 Objectives

After going through this unit, you will be able to know and draw

(i) Histogram

(ii) Frequency polygon

(iii) Frequency curve

(iv) Ogives

3.3 Graphic Representation of Frequency Distribution

Graphic Representation of Frequency Distribution is a powerful tool of data presentation and interpretation because the shape of graph provides easy answers to several important questions. Normally the frequency distribution, as a tabular representation is not able to highlight the essential characteristics of the data as apparently as its graphic presentation may do. The shape of graph offers an exact idea of the variations, its skewness, peakedness, modes, extremes, outliers, spread etc. inherent in the distribution of data. Accordingly frequency

distribution graphs serve as effective tools of a quick analysis and effective comparison between two or more distributions. The pattern of variations and the points of contrast become quite obvious when the graph of one frequency distribution is superimpose on the other. Following are important methods of graphical representation of Frequency Distributions.

3.4 Histogram

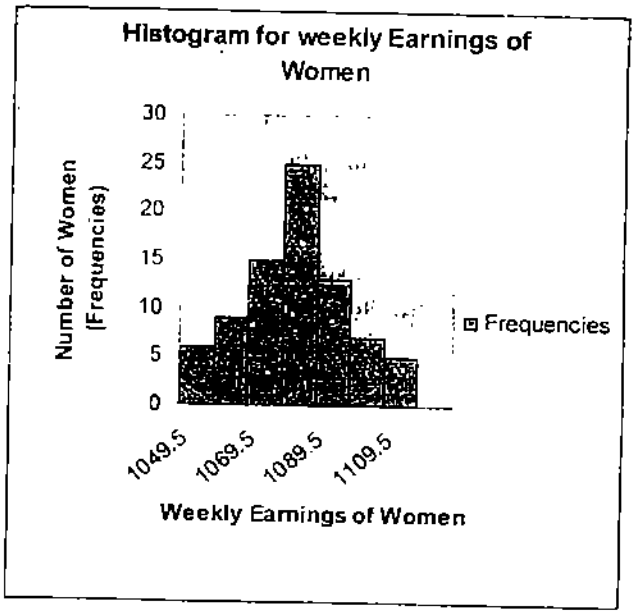
A histogram is the simplest and common form of graphical representation of Frequency Distributions. It groups the values of a variable into evenly spaced groups or intervals and plots a count of the number of cases (frequency) in each group. The count can also be expressed as a percentage. The purpose of a histogram is to graphically summarize the frequency distribution of a univariate distribution. Along the vertical Y-axis, which always begins with zero at the point of origin, are measured the class frequency. Rectangular bars are then raised over successive class intervals with their base equal in width on the X-axis. The height of each bar measured on Y-axis is kept equal to the corresponding class frequency. The area of the bar corresponding to each class interval is given by its class frequency "f' multiplied by the width of class interval C. Since frequency distributions may have equal or unequal class intervals, the procedure for drawing a histogram is described separately for both the situations as under.

(i) Histogram for Equal Intervals

Figure 3.1 represents the histogram of the frequency distribution with equal interval given in Unit 2. The horizontal X-axis is divided by marking dots into equal parts numbering two or three more than the number of class intervals comprising the distribution. Starting from left

not necessarily with zero. each dot is labeled by the lower class limit of the successive class, leaving a space equal to the size of one class interval on the either extreme side. At times, the horizontal scale is also used to show the mid points of the successive class intervals.

Figure 3.1 showing Histogram for data given in table 1.2.4



(ii) Histogram for Unequal Class Intervals

distribution with unequal class intervals is materially not different. It requires only minor adjustments in the spacing dots marked on the X-axis. Here frequency densities are plotted against class intervals instead of frequencies.

$$\text{Frequency density of any class} = \frac{Class\ frequency}{Width\ of\ the\ class}$$

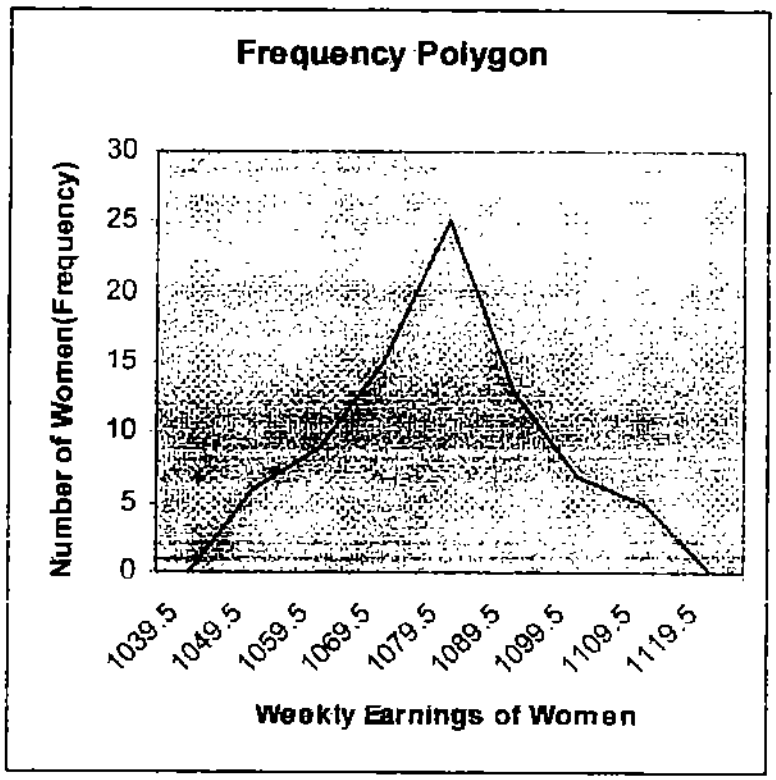
The method of calculating frequency density is simple and they are obtained by dividing the frequencies of each class by their respective

class widths. A histogram for an open ended distribution is drawn essentially the same way, except that the open end class are not considered. Limits are chosen arbitrarily so that the width of the class intervals become equal to the preceeding (or succeeding) class.

3.5 Frequency Polygon

Frequency Polygon represents yet another way of depicting a frequency distribution in the form of a graph. Given the histogram in figure, a Frequency Polygon is drawn by making dots at the mid-points of the top of the each bar and joining them by means of straight lines. The polygon so obtained is closed at the end by joining the top base mid points of the first and the last rectangles with the mid points of the next outlying interval on either side. The mid-points of these two outlying intervals fall on their bottom base, meaning zero class frequencies.

Figure showing frequency polygon



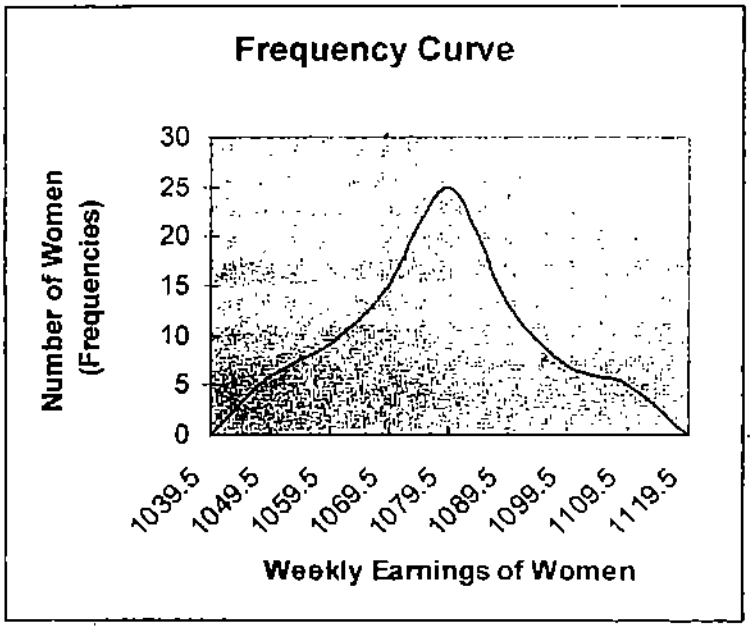
In fact constructing a frequency polygon does not necessarily erequire a histogram being drawn frist. It can be obtained directly by plotting dots above each interval mid point at heights equal to the corresponding class frequencies and joining them by mean of straight lines. The polygon is closed on either side exactly the same way as explained above. However, the X-axis measures the successive class mid points, and not the lower class limits. This is shown in above Figure.

3.6 Frequencies Curve

Frequency curves are obtained by smoothing the frequency polygon drawing a free hand smooth curve through the various points that yield a frequency polygon on joining. Serious limitation of a smooth curve drawn in free hand is that no two persons will ever smooth the polygon in exactly the same way. Despite of this limitation, the need for smoothing a polygon can not be overemphasized. A frequency polygon does not get uneven so much owing to the inherent irregulaties in the data. Instead it becomes more erratic on account of selection of class width which makes the class frequencies change abruptly. The real advantage of smoothing thus lies in eliminating the abrupt behavior of the polygon and making it more representative of the true variations in the data.

It may be noticed that a frequency distribution based on a larger number of sample data observations will have a smoother frequency polygon, It will closely approximate a polygon based on the entire population as the number of observations comprising the sampe increases.

Figure showing frequency curve



A polygon may assume a variety of shapes, more frequently encountered among them are either symmetrical or skewed in shape. Some others not so common are J-shaped, V-shaped, S-shaped, bimodal, etc.Class frequency are measured along the vertical Y-axis, which always begins with zero at the point of origin. Rectangular bars are then raised for successive class intervals with their base equal in width on the X-axis. The height of each bar measured on Y-axis is kept equal to the corresponding class frequency. The area of the bar corresponding to each class interval is given by its class frequency multiplied by the width of the class intervals.

3.7 Cumulative Frequency Curves or Ogives

A cumulative frequency curve, popularly known as Ogive, is another form of graphic presentation of frequency distribution .As an illustration consider the frequency distribution presented in Unit 2. The first step in drawing a cumulative frequency curve is to obtain

cumulative frequencies denoted as 'cf and record them in separate column. Cumulative frequencies can be of less than type or more than type. Cumulative frequencies less than type are obtained by adding successive frequencies from top to bottom and 'more than type' cumulative frequencies are obtained by adding successive frequencies from bottom to top.

Procedure for drawing a cumulative frequency curve, or an ogive, is same as any frequency curve. The only difference being that the Y-axis now shows the total cumulative frequencies. The X-axis is labeled with the upper limits in the case of "less than type" ogive and lower limits in the case of 'more than type' ogive.If both less than type and more than type ogive are plotted on the same graph they intersect at median of observations. The ogives can also be smoothed by free hand as frequency polygon.

Figure : Cumulative frequency curve or ogive (More than type)

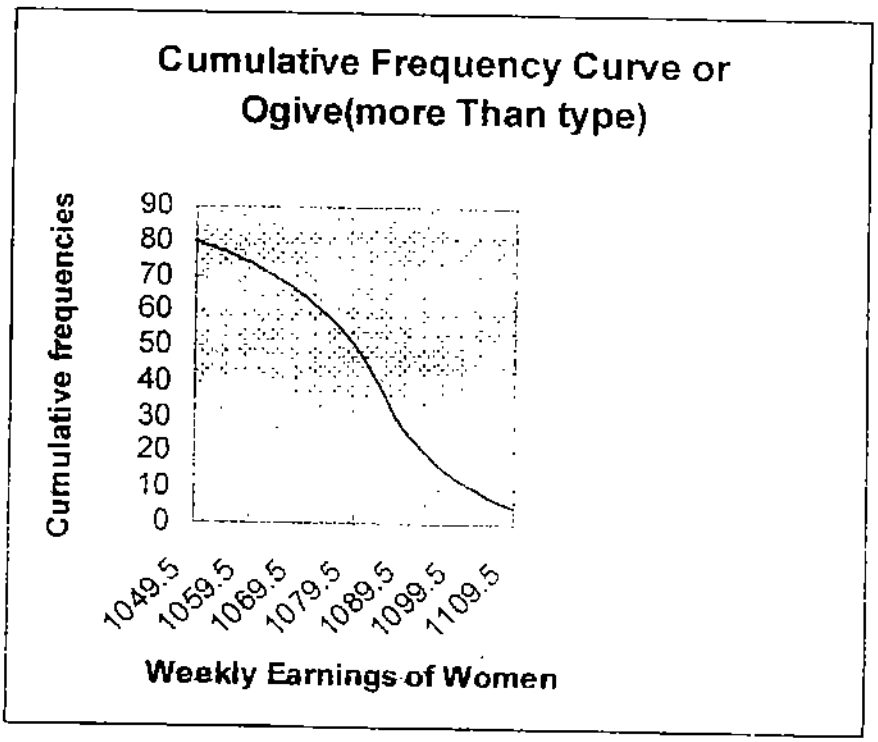
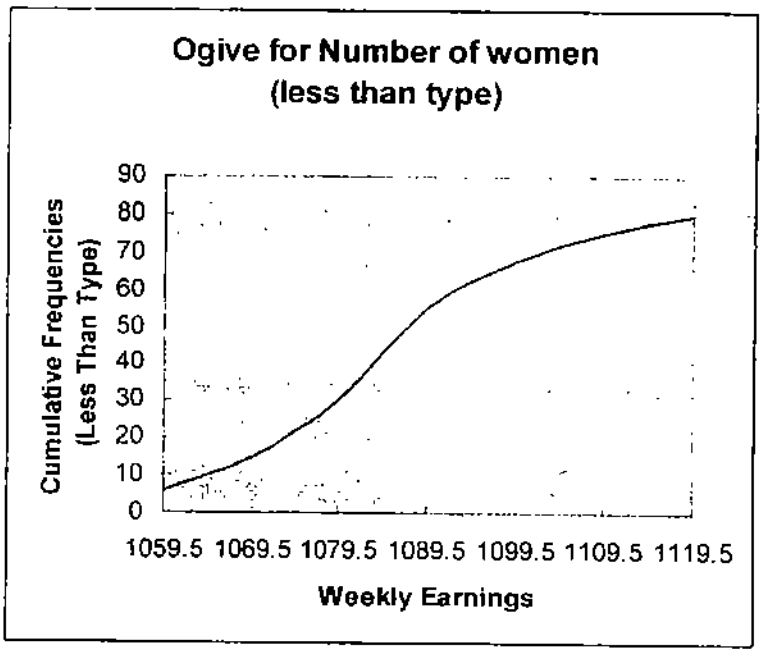


Figure : Cumulative frequency curve or ogive (less than type)



3.8 Summary

Geogrphical Representation of frequency distribution is a powerful tool of data representation and interpretation because the shape of graph provides easy answers to several important questions. A histogram is the simplest and common form of graphical representation of data. Frequently polygon and frequency curves are other important tools of graphical representation . A polygon may assume a variety of shapes, more frequently encountered among them are either symmetrical or skewed in shape. Ogives are another form of graphic representation of frequency distribution. The point of intersection of two ogives gives the median.

In a histogram the areas of the rectangles equals the corresponding frequencies whereas in bar diagrams height of the bars equals the frequency.

3.9 Self-Assessment Exercises

P-1 Histogram and Historigram are the :

(a) same (b) different

P-2 False base line is used in :

(a) Higstogram (b) Historigram (c) both

P-3 Use of false base line is :

(a) must (b) desirable (c) unwanted

P-4 Between two rectangles of a hisrogram a gap is :

(a) necessary (b) allowed (c) never allowed

P-5 In a histogram, if the width of a class is doubled that that of other classes, then its frequency is :

(a) doubled (b) halved (c) no change

P-6 Class limits and class boundaries are :

(a) Always same (b) Always different (c) not always different.

P-7 A time series data is presented by means of :

(a) histogram (b) histotrigram (c) bar diagram

(d) ogive

P-8 The two types of ogives cuts each other at :

(a) Median (b) Q_1 (c) Q_3 (d) Mean

P-9 A historigram is a :

(a) diagram (b) graph (c) table (d) text

P-10 Ogive curve occurs for :

(a) more than type distribution

(b) Less than type distribution

(c) both (a) and (b)

(d) none of (a) and (b)

P-11 Ogive for more than type and less than type distribution intersect at :

(a) mean (b) median (c) mode (d) origin

P-12 In case of frequency distribution with classes of unequal widths, the heights of bars of a histogram are proportional to :

(a) class frequency (b) class intervals

(c) Frequencies in percentage (d) Frequency densities

P-13 In a histogram with equal class intervals, the height of rectangles are proportional to :

(a) mid-values of the classes respective (b) frequencies of the classes

(c) either (a) or (b) (d) neither (a) nor (b)

3.10 Further Readings

(1) Goon, A.M., Gupta, M.K., Dasgupta, B: Fundamentals of Statistics. Vol. One, The World Press Pvt. Ltd., Calcutta

(2) Yule, G.U. and Kandall, M.G.: An Introduction to the theory of Statistics. Charles Griffin & Company Limited

(3) Weatherburn, C.E: Mahematical Statistics.

NOTES

U.P. Rajarshi Tandon Open
University, Allahabad

UGSTAT-01
STATISTICAL METHODS

Block - II

Measures of Central Tendency and Dispersion

Introduction

This is the second block on statistical methods. There are two units in this block.

Unit-1 of this block deals with measures of central tendency. Once data have been collected and represented, one may like to know the particular value around which the data has the tendency to concentrate. This value is known as measure of central tendency. Various measures of central tendency along with their characteristics have been discussed in this unit.

Unit – 2 deals with measures of dispersion. Once data have been represented and a measure of central tendency has been located, one may like to know the scatterness of the data around this measure of central tendency. Various measures of dispersions have been defined and their characteristics have also been discussed.

Unit-1 Measures of Central Tendency

Structure

1.1 Introduction

Statistical methodology is a comprehensive term which includes almost all the methods involved in the colletion, processing, condensing and analysing of data. The data collected from the field for a number of items vary greatly in their qualitative as well as quantitative nature. For example, the rainfall at a particular region is erratic in nature and shows variation from year to year, month to month and even day to day. The condensation of data in terms of maps, charts, diagrams, etc. is a first and necessary step in rendering a long series of observations comprehensible. But for practical purposes it is not enough, particularly when we want to compare two or more different series of data, e.g. we may wish to compare the distribution of status in two races of man, or the birth rates in India in two successive decades, or the rainfall in two different regions, or the number of wealthy people in two different countries. For such problems, there are certain statistical techniques one of which is a measure of central tendency.

It is found that the observations have a tendency to cluster round a central value, and this characteristic of observations is called the **'central tendency'**. Any statistical measure which gives the point round which the observations have a tendeny to cluster is known as a **'measure of central tendency'**. The central value of the variable in any series of obsevations is useful in finding the location of the distribution and so it is also called an average. Thus, an **'average'** of a series of measurements is a single value of the variable which is a satisfactory representative of the distribution.

In this unit have been highlighted different measures of central tendency are covered. Various situations where they find calculation of these measures for ungrouped and grouped data are described.

1.2 Objectives

After studying this unit you will be able to

- Understand the meaning of central tendency of data,
- Compute common measures of central tendency, i.e., mean, median and mode.

- Compute the various measures of partitions of data such as quartiles, deciles and percentiles.
- Understand how to choose proper measure of central tendency. .

1.3 Arithmetic Mean (Ungrouped Data)

The arithmetic mean of a series of n observations X_1, X_2, X_3,.... X_n is obtained by summing up the values of all the observations and dividing the total by the number of observations. Thus,

$$\bar{X} = \frac{Sum\ of\ the\ observations\ (or\ values)}{Number\ of\ observations}$$

$$= \frac{X_1 + X_2 + X_3 + ... + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Where $\sum$ (sigma) stands for summation and X_i is the ith value of the observation (variable).

Example 1.1

The rainfall records in a month of 10 regions of a State is given below. Compute the average rainfall of the month for the State.

Region :	1	2	3	4	5	6	7	8	9	10
Rainfall : (in mm)	17.6	10.1	11.4	18.5	10.5	14.3	8.9	13.4	10.6	12.5

Solution

$$Mean\ (\bar{X}) = \frac{\sum_{i=1}^{10} X_i}{10}$$

$$= \frac{X_1 + X_2 + ... + X_{10}}{10}$$

$$= \frac{17.6 + 10.1 + 11.4 + ... + 10.6 + 12.5}{10}$$

$$= \frac{127.8}{10} = 12.78 mm$$

The computation of Arithmetic mean using short cut method is discuss below:

1.3.1 Short- Cut Method

This method is applied to avoid lengthy calculations. When the individual set of readings are large in size, an arbitrary value is selected as a working mean (known as assumed mean) and differences between the working mean and the individual readings (known as deviations from the assumed mean) are worked out. By summing these differences and dividing by the number of readings we can get the mean of deviations from assumed mean. Let $X_1, X_2, X_3 \ldots\ldots X_n$ be n individual readings on the variable and let A be the working mean. Let $d_1, d_2, d_3 \ldots\ldots d_n$ denote the differences between the working mean and individual values $X_1, X_2, X_3 \ldots\ldots X_n$ respectively. Then mean $\bar{X}$ of X, in terms of the mean $\bar{d}$ of differences, is calculated as :

$$\bar{d} = \frac{d_1 + d_2 + d_3 \ldots + d_n}{n}$$

$$= \frac{(X_1 - A) + (X_2 - A) + (X_3 - A) + \ldots + (X_n - A)}{n}$$

$$= \frac{\sum_{i=1}^{n} X_i - nA}{n} = \frac{\sum_{i=1}^{n} X_i}{n} - A$$

$$= \bar{X} - A$$

or,

$$\bar{X} = A + \bar{d} = A + \frac{1}{n}\sum_{i=1}^{n} d_i$$

True mean = guessed mean + (sum of deviations from guessed mean / number of cases)

This is useful, if the size of frequencies are large. An illustration is given below.

Example 1.2

Calculate the mean for the following scores : 60, 65, 74, 85, 95.

Solution:

Table 2.1 Distribution of Scores

Xi (Scores)	Xi – 74
60	–14
65	–9
74	0
85	+11
95	+21
	+9

Then, mean score is

$\bar{X} = 74 + (+9/5) = 74 + 1.8 = 75.8$

1.3.2 Grouped data (Discrete frequency distribution)

In a discrete series, let the individual readings $X_1, X_2, \ldots X_n$ of the variable X occur (have frequencies) $f_1, f_2, \ldots, f_n$ times respectively. The product Xi. fi is the sum of all X_i's in the data and $\sum_{i=1}^{n} X_i f_i$ is the sum of all the observations. Then the mean of X is obtained by summing the product of individual readings with corresponding frequencies and dividing the total by the sum of frequencies, i.e.

Mean $\bar{X} = \dfrac{X_1 f_1 + X_2 f_2 + \ldots + X_n f_n}{f_1 + f_2 + \ldots + f_n}$

$$= \frac{\sum_{i=1}^{n} X_i f_i}{\sum_{i=1}^{n} f_i} \quad or \quad \frac{\sum_{i=1}^{n} X_i f_i}{N}$$

where $N = \sum_{i=1}^{n} f_i$ is the total number of observations.

The various steps involved in the calculation of $\bar{X}$ under this method are:

(i) Multiply the frequency of each row with the concerned variable $(X.f)$ and total them. It will be ΣXf.

(ii) Add up all the frequencies (Σf).

(iii) Divide ΣXf. by Σf or N to find out the arithmetic mean $(\bar{X})$.

Short cut method :

To find our arithmetic mean in a discrete series by this method the following steps are taken :

(i) Any value of the distribution may be taken as working mean or arbitrary mean, say A (preferably it should be near the middle of the frequency distribution).

(ii) Take deviation d_x of the variable X from the working mean (X-A) and denote it by dx.

(iii) Multiply each d_x by its respective f and denote it by fd_x.

(iv) The arithmetic mean $(\bar{X})$ is then calculated with the help of following formula :

$$\text{Mean} = \bar{X} = A + \frac{1}{N}\sum_{i=1}^{n} f_i d_{x_i}$$

Example 1.3

Below are given the number of children born per family in 735 families in a locality. Calculate the average number of children born per family in the locality.

Table – 1.2 :

Number of Children Born per Family

Number of childern born per family	Number of families	Number of children born per family	Number of families
0	96	7	20
1	108	8	11
2	154	9	6
3	126	10	5
4	95	11	5
5	62	12	1
6	45	13	1

Solution:

Computation of the average number of children born per family :

Table 1.3

Showing the calculation

No. of children born per family *(X)*	No. of families *(f)*	*X.f* *(Xf)*
0	96	0 x 96 = 0
1	108	1 x 108 = 108
2	154	2 x 154 = 308
3	126	3 x 126 = 378
4	96	4 x 95 = 380
5	62	5 x 62 = 310
6	45	6 x 45 = 270
7	20	7 x 20 = 140
8	11	8 x 11 = 88
9	6	9 x 6 = 54
10	5	10 x 5 = 50
11	5	11 x 5 = 55
12	1	12 x 1 = 12
13	1	13 x 1 = 13
Total	735	2166

Here $N = \Sigma f = 1343$; $\Sigma Xf = 2166$

Average Number of Children Born per Family is given by

Mean $(\overline{X}) = \frac{\Sigma Xf}{\Sigma f} = \frac{2166}{735} = 2.9$ children.

Example 1.4

The following table gives the distribution of units under different heights in a certain region. Compute the mean height of the region :

Table 1.4

Height of units

Height (in metre) :	200	600	1000	1400	1800	2200
Number of Units :	142	265	560	271	89	16

Solution:

(a) By direct method

Height (in metre) *(X)*	No. of Units *(f)*	*X.f* *(Xf)*
200	142	200 x 142 = 28400
600	256	600 x 256 = 159000
1000	560	1000 x 560 = 560000
1400	271	1400 x 271 = 379400
1800	89	1880 x 89 = 160200
2200	16	2200 x 16 = 35200
Total	1343	1322200

Here $N = \Sigma f = 1343;\ \Sigma Xf = 1322200$

Mean height $(\bar{X}) = \frac{\Sigma Xf}{\Sigma f} = \frac{1322200}{1343}$

$= 984.51$ metres

(b) By short- cut method

Let the working mean *(A)* = 1400

Table 1.5

Distribution of Height of Units

Height (in Metres *(X)*	$X - A$ *(d)*	No. of units *(f)*	*d.f* *(df)*
200	—1200	142	—1200 x 142 = 170400
600	—800	265	—800 x 265 = 212000
1400 A	0	271	0 x 271 = 0
1800	+ 400	89	+400 x 89 = 35600
2200	+ 800	16	+ 800 x 16 = 12800
Total		1342	—558000

Here $N = \Sigma f = 1343$; $\Sigma fd = -558000$

Mean height $(\bar{X}) = A + \frac{\Sigma fd}{N}$

$= 1400 + \frac{(-558000)}{1343} = 984.51$ metres.

1.3.3 Grouped Data (Continuous Frequency Distribution)

When the measurements are given in the grouped form, the mean is computed by multiplying the various mid-values m_i with their respective frequencies f_i where i =1,2...,n and dividing the product total by the sum of frequencies or total number of observationss. If $m_1, m_2, \ldots, m_n$ are the mid-values or class marks corresponding to frequencies $f_1, f_2, \ldots, f_n$ then the mean is

$$\bar{X} = \frac{m_1 + m_2 f_2 + \ldots\ldots\ldots + m_n f_n}{f_1 + f_2 + \ldots\ldots + f_n}$$

$$= \frac{\sum_{i=1}^{n} m_i f_i}{\sum_{i=1}^{n} f_i}$$

$$= \frac{\sum_{i=1}^{n} m_i f_i}{N}$$

where n stand for the number of groups and N denotes the total number of observations.

Short-cut method : If the class - groups formed by individual readings are large, the following procedure is adopted for computing arithmetic mean in a continuous series as follows :

(i) Find out mid-values of the class intervals, and assume one of the mid-values (preferably in the middle of the distribution) as working mean *(A).*

(ii) Calculate deviations of the mid-values from the working mean (m-A) and denotes by d. .

(iii) Multiply each d by its respective frequency to find out fd.

(iv) Calculate arithmetic mean $\bar{X}$ by the following formula :

$$\text{Mean } (\bar{X}) = A + \frac{\sum_{i=1}^{n} f_i d_i}{N}$$

Step-Deviation Method : Further if the deviations are large and intervals among consecutive mid-values are equal, deviations can be made small by dividing them by the class width, i.e.. if $d_i' = \frac{(m_i - A)}{h}$, the formula for calculating arithmetic mean by this method is

$$\text{Mean } (\bar{X}) = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} \times h$$

where h is the size of the class intervals.

Example 1.5

The rainfall of 66 districts in a particular year is given below. Computed the average annual rainfall.

Rainfall :
(in inches) 0—10 10—20 20—30 30—40 40—50 50—60

No. of Districts :	22	10	8	15	5	6

Solution :

(a) By direct method

Table 1.6

Rainfall in Districts in Particular Year

Rainfall (in inches) *(X)*	Mid-values *(m)*	No. of districts *(f)*	*m.f* *(mf)*
0—10	$\frac{0+10}{2} = 5$	22	5 x 22 = 110
10—20	$\frac{10+20}{2} = 15$	10	15 x 10 = 150
20—30	$\frac{20+30}{2} = 25$	8	25 x 8 = 200
30 — 40	$\frac{30+40}{2} = 35$	15	35 x 15 = 525
40 — 50	$\frac{40+50}{2} = 45$	5	45 x 5 = 225
50 — 60	$\frac{50+60}{2} = 55$	6	55 x 6 = 330
Total		66	1540

Here $N = \Sigma mf = 1540$

$$\text{Mean } (\bar{X}) = \frac{\Sigma mf}{\Sigma f} = \frac{1540}{66}$$

$= 23.33$ inches

(b) Short Cut Method

Rainfall in inches *(X)*	Mid-values *(m)*	No. of districts *(f)*	No. of districts *(f)*	*d* x *f* *(df)*
0—10	5	—30	22	— 30 x 22 = — 660
10—20	15	—20	10	—20 x 10 = — 200
20 — 30	25	— 10	8	— 10 x 8 = —80
30 — 40	35(A)	0	15	0 x 15 = 0
40 — 50	45	10	5	10 x 5 = 50

50 — 60	55	20	6	20 x 6 = 120
Total			66	—770

Here $N = \Sigma f = 66$; $\Sigma df = -770$

Mean $(\bar{X}) = A + \dfrac{\Sigma fd}{N}$

$= 35 + \dfrac{-770}{66} = 23.33$ inches

(c) By Step - deviation Method

From the given data $h = 10$

Rainfall in inches (X)	Mid-values (m)	(d')	No. of districs (f)	d x f (df)
0—10	5	$\dfrac{5-35}{10} = -3$	22	— 3 x 22 = — 66
10—20	15	$\dfrac{15-35}{10} = -2$	10	— 2 x 10 = —20
20—30	25	$\dfrac{25-35}{10} = -1$	8	— 1 x 8 = — 8
30—40	35(A)	$\dfrac{35-35}{10} = 0$	15	0 x 15 = 0
40—50	45	$\dfrac{45-35}{10} = +1$	5	1 x 5 = 5
50—60	55	$\dfrac{55-35}{10} = +2$	6	2 x 6 = 12
Total			66	—77

Here $N = \Sigma f = 66$; $\Sigma fd' = -77$

Mean $(\bar{X}) = A + \dfrac{\Sigma fd'}{N} \times h$

$= 35 + \dfrac{(-77)}{66} \times 10 = 23.33$ Inches

Some more Examples

Example 1.6

The amounts of money, in thousand dollars, that a sample of people contributed to political campaigns in an election are : 1.2, 5, 25, 10, 0, 2, 0, 5, 10. Answer the following questions :

(i) What is the total amount of money contributed ΣX_i ?

(ii) What is the money contributed by seventh person, X_7?

(iii) What is the sample size ?

(iv) Find the mean and interpret it.

Solution :

First make a frequency distribution as shown below.

x_i	f_i	$f_i x_i$
0	2	0
1	1	1
2	2	4
5	2	10
10	2	20
25	1	25
Total	10	60

(i) Total contribution, $\Sigma f_i X_i = 0 + 1 + 4 + 10 + 20 + 25 =$ \$ 60 thousand

(ii) Contribution of seventh person, $X_7 =$ \$2 thousand

(iii) Sample size – number of case = n = 10

(iv) Mean $\overline{X} = \Sigma f_i X_i / n = 60 / 10 = 6$ thousand.

This value of mean shows that the average contribution to political campaigns in the election was \$6,000. This also tells that most of the contributions are located around a value of \$6,000. Thus, the central of the

distribution of campaign contributions has been located, It is not yet how well this quantity measures average.

Example 1.7

Obtain an estimate of the value of missing frequency? if the mean of the distribution is 22.5.

C.I.	f.	x (Mid value)	fx
0 - 10	6	5	30
10 - 20	9	15	135
20 - 30	a	25	25a
30 - 40	10	30	300
40 - 50	3	45	135
	40		600 + 25a

Let a be the unknown frequency.

$$\bar{x} = \frac{1}{N}\sum f(x)$$

$$22.5 = \frac{1}{40} \times (600 + 25a)$$

$$22.5 = \frac{25(24 + a)}{40}$$

22.5 x 40 - 25 (24 + a)

900 = 25 (24 + a)

900 / 25 = 24 + a

36 = 24 + a

a = 36 - 24 = 12. a= 12.

The missing value is 12.

Example 1.8 (On wrongly taken value).

C.I.	f.	x Mid value	fx
0 - 10	5	5	25
10 - 20	12	15	180
20 - 30	18 (15)	25	450 (375)
30 - 40	11	35	385
40 - 50	3	45	135
Total	49		1175

$$\bar{x} = \frac{1175}{49} = 23.98$$

Since the value 18 is wrong and the 15 is correct value than the calculations are as followed:

Present $N = \sum f_i = 49.$

Then correct N is

(Corrected) N = 49 - 18 + 15 = 46.

Similarly.

(Corrected) $\Sigma fx = 1175 - 450 + 375 = 1100$

Then the corrected mean is.

(Corrected) $\bar{x} = \frac{1100}{46} = 23.91$

1.3.4 Properties of Arithmetic Mean or Fundamental Theorems On Arithmetic Mean

1. First property of mean :

The sum of the deviations about the arithmetic mean equals zero. Mathematically.

$$\Sigma\left[f_i\left(X_i - \bar{X}\right)\right] = 0$$

Proof: $\Sigma f_i(X_i - \bar{X}) = \sum f_i X_i - \sum f_i \bar{X}$

$$= \sum_i f_i X_i - \bar{X} \sum_i f_i \quad = (Since\ \bar{X}\ be\ the\ independent\ of\ i)$$

$$= N\bar{X} - \bar{X}.N = 0$$

Since,

$$\left(\bar{X} = \frac{1}{N} \sum_i f_i X_i \quad where\ N = \sum_i f_i \ then \sum_i f_i X_i = N\bar{X} \right)$$

Hence proved.

This property says that if the mean is substracted from each score, the sum of the differences will equal zero. This property results from the fact that the mean is the balance point of the distribution. The mean can be thought of as the fulcrum of a seesaw. When the scores are distribution along the seesaw according to their values, the mean of the distribution occupies the position where the scores are in balance. This is known as first property of mean.

2. Second property of mean :

The sum of the squared deviations of all the scores about their arithmetic mean is minimum . That is,

$$\Sigma \left[f_i \left(X_i - \bar{X} \right)^2 \right] = \text{minimum}$$

This is an important characteristic and is used in many areas of statistics. particularly in regression analysis.

Proof : Let us suppose that the sum of square of deviations from point (a).

$$\sum f_i (X_i - a)^2 = K$$

According to the principle of maxima, K will be minimum if,

$$\frac{\partial k}{\partial a} = 0 \ and \ \frac{\partial^2 k}{\partial a^2} > 0.$$

Now $\frac{\partial k}{\partial a} = (-2) \Sigma f_i (X - a) = 0$

$$\Rightarrow \sum f_i (x_i - a) = 0 \quad \Rightarrow \sum f_i X_i - Na = 0 \Rightarrow a = \frac{1}{N} \sum f_i X_i = \bar{X}$$

Again, $\frac{\partial^2 k}{\partial a^2} = (-2)\sum f_i(-1) = 2\sum f_i = 2N > 0$

Hence, K is minimum at a = X and,

$$\sum\left[f_i\left(X_i - \bar{X}\right)^2\right] = \text{minimum}$$

Remarks, We shall see later on that,

$$\sigma^2 = \frac{1}{N}\sum f_i(X_i - \bar{X})^2$$

Is a measure of dispersion

3. Combined property of mean.

If $\bar{X}_1$ and $\bar{X}_2$ be the means of two series of sizes n_1 and n_2 respectively, then the mean $\bar{X}$ of the combined series can be computed as :

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

Proof. : If $\bar{X}_1$. be the mean of series $X_{11}\ X_{12}\ldots\ldots X_{1n_1}$

and $\bar{X}_2$ be the mean of series $X_{21}\ X_{22}\ldots\ldots X_{2n_2}$

Then, by definition,

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i = \frac{1}{n_1}\left(X_{11} + X_{12} + \ldots\ldots + X_1 n_1\right)$$

$$\bar{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_j = \frac{1}{n_2}\left(X_{21} + X_{22}\ldots\ldots + X_{2n_2}\right)$$

The combined series is $X_{11}\ X_{12}\ldots\ldots X_{1n_1} X_{21} X_{22}\ldots\ldots X_{2n_2}$

The mean is

$$\bar{X} = \frac{1}{n_1 + n_2}\left[\left(X_{11} + X_{12} + \ldots\ldots + X_{1n_1}\right) + \left(X_{21} + \ldots\ldots + X_{2n_2}\right)\right]$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} X_i + \sum_{j=1}^{n_2} X_j\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1\overline{X}_1 + n_2\overline{X}_2\right]$$

Similarly, if $\overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots \ldots \overline{X}_k$ be the means of k series of sizes $n_1, n_2, \ldots \ldots, n_k$ respectively, then the mean $\overline{X}$ of combined series is

$$\overline{X} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + \ldots + n_n\overline{X}_n}{n_1 + n_2 + \ldots + n_n}$$

Example 1.9

The average ages of 250 males and 210 females in a village are 41.6 and 38.5 years, respectively. Find the average age combining both males and females together.

Solution :

Here (Combined average) is $N_1 = 250, \overline{X}_1 = 41.6$ *years* and $N_2 = 210, \overline{X}_2 = 38.5$ *Years*

Therefore,

$$\overline{X} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

$$= \frac{250 \times 41.6 + 210 \times 38.5}{250 + 210} = \frac{10400 + 8085}{460}$$

$= 40.18$ years.

1.3.5 Advantages of Mean

(i) The mean is sensitive to the exact values of all the scores in the distribution. Since you have to add all the scores to calculate the mean, a change in any of the scores will cause a change in the mean.

(ii) Mean is very sensitive to extreme scores. If we add an extreme score (one that is very far from the mean), it would greatly disrupt the balance. The mean would have to shift a considerable distance to reestablish balance.

The mean is more sensitive to extreme than is the median or the mode.

This is known as 2nd property of mean.

(iii) Of the measures used for central tendency, the mean is least subject to sampling variation under most circumstances. If repeated samples are drawn from a population, the mean would vary from sample to sample. The same is true for the median and the mode. However, the mean varies less than these other measures of central tendency. This is very important in inferential statistics and is a major reason why the mean is used in inferential statistics whenever possible.

(iv) It takes into account all the scores in a distribution So, mean offers a good representation of the central tendency by making use of the most information.

(v) Mean is used in many statistical formulas, making it a more widely used measure.

1.3.5 Limitations of Mean

Mean can be misleading if there are extreme values in the distribution, for example, if the distribution is skewed (asymmetrical) or the level of measurement is less than interval. Sometimes people are interested in misleading others by making use of 'illegitimate' statistics. The following example illustrates this point.

Example 1.10

The amounts of money that a sample of people contributed to political campaigns in last election were, in thousand rupees : 1, 2, 5, 25, 10, 0, 2, 0, 5, 10, 500. Calculate the mean.

Solution :

Make a table that contains columns: X_i, f_i and f_i Sum all the entries in columns X_i f_i and f_i to get :

$f_i X_i = 560$ ' $f_i = n = 11$

Mean, $\overline{X} = f_i X_i / n = (1 + 2 + 5 + 25 + \ldots\ldots + \ldots\ldots + 10 + 500)/10 = 560/11$

$= 50.91$ or Rs. 50.910.

A mean of Rs.50.91 thousand suggests that the typical contribution was Rs. 50,910. We notice that the mean in this example is not at all a measure of central tendency. All but one person contributed less than the mean. This is due to the presence of an extreme contributor who contributed Rs. 500,000. This high value inflates the mean making it a misleading statistics in each situations.

1.4 Geometric Mean

In case of finding the average of rates and ratios, geometric mean is more useful measure than others, e.g., in finding the population increase, simple and compound interests, etc.

Case 1 : In case of ungrouped data, it is obtained by multiplying together all the values of the variable and extracting the relevant root of the product, i.e .if $X_1, X_2, \ldots X_n$ are n values of a variable under study, then the geometric mean (G.M.) is computed as :

$$\text{G.M.} = (X_1, X_2, X_3 \ldots X_n)^{\frac{1}{n}} \quad ; \quad (X_i > 0)$$

To facilitate the computation, one can make the use of logarithms as :

$$\log(\text{G.M.}) = \frac{1}{n} log\,(X_1, X_2, X_3 \ldots, X_n)$$

$$= \frac{1}{n}[\log X_1 + \log X_2 + \log X_3 \ldots + \log X_n]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} \log X_i\right]$$

$$\text{So G.M.} = \text{Antilog} \quad \frac{1}{n}\left[\sum_{i=1}^{n} \log X_i\right]$$

Thus the logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of individual measurements.

Case 2 : In case of grouped data, if $f_1, f_2, \ldots, f_n$ be the frequencies corresponding to the individual values $X_1, X_2, \ldots X_n$ then G.M. is computed as :

$$\text{G.M.} = \left(X_1^{f_1}, X_2^{f_2}, X_3^{f_3} \ldots X_n^{f_n}\right)^{\left[\frac{1}{\sum f_i}\right]} \quad ; \quad (X_i > 0)$$

$$\text{or } \log \text{G.M.} = \frac{1}{\sum_{i=1}^{n} f_i}\left[f_1 \log X_1 + f_2 \log X_2 + \ldots + f_n \log X_n\right]$$

$$= \frac{1}{\sum_{i=1}^{n} f_i}\left[\sum_{i=1}^{n} f_i \log X_i\right]$$

$$\text{So G.M.} = \text{Antilog}\left\{\frac{1}{\sum_{i=1}^{n} f_i}\left[\sum_{i=1}^{n} f_i \log X_i\right]\right\}$$

Here, log G is the weighted mean of $\log_G y_i$, s with weights $f_1, f_2, \ldots\ldots f_n$.

Additive property of Geometric Mean.

If G_1 and G_2 are the geometric means of two series with respective sizes n_1 and n_2, the Combined Geometric mean G is

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

Proof

Suppose that $X_{11}, X_{12} \ldots\ldots\ldots X_{1n}$ be the first series and $X_{21}, X_{22} \ldots\ldots X_{2n_2}$ be the 2nd series with sizes n_1 and n_2. Then

$$G_1 = \left[\prod_{i=1}^{n_1} x_{1_i}^{1}\right]^{1/n_1} \text{ and } G_2 = \left[\prod_{j=1}^{n_2} x_{2_j}\right]^{1/2_n}$$

$$\log G_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \log_e x_{1_i}, \quad \log G_2 = \frac{1}{n_2} \log \sum_{j=1}^{n_2} \log_e x_{2_j}$$

G.M. of combined series $X_{11}\, X_{12} \ldots\ldots X_{1n_1}, X_{21} X_{22}, \ldots\ldots X_{2n_2}$ is,

$$G = \prod_{k=1}^{n_1+n_2} x_k^{\frac{1}{n_1+n_2}} \Rightarrow \log G = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1} \log x_{1_i}^{n_1} + \sum_{j=1}^{n_2} \log x_{2_j}^{n_2}\right]$$

$$\Rightarrow \log_e G = \frac{1}{n_1+n_2}\left[n_1 \log_e G_1 + n_2 \log_e G_2\right]$$

Proved.

Example 1.11

The monthly average temperature of a station for five months are given as 16.2, 23.4, 20.6, 33.4 and 16.4 degree centigade. Find the mean temperature of the station.

Solution :

$$\text{G. M.} = (16.2 \times 23.4 \times 20.6 \times 33.4 \times 16.4)^{\frac{1}{5}}$$

or

$$\log \text{G.M.} = \frac{1}{5}[\log 16.2 + \ldots\ldots + \log 16.4]$$

$$= \frac{1}{5}[1.2095 + 1.3692 + 1.3139 + 1.5238 + 1.2148]$$

$$= \frac{1}{5} \times 6.6312 = 1.3262$$

sc , G.M. = Antilog [1.3262] = 21.1934 degree centigrade.

Example 1.12

From the following data, calculate the G.M.

Class group :	0—10	10—20	20—30	30—40	40—50
No. of observations :	14	23	27	21	15

Calculation :

Rainfall in inches	Mid-values *(x)*	Log *(x)*	No. of observations *(f)*	$f \log x$
0—10	5	0.69897	14	9.78558
10—20	15	1.17609	23	27.05007
20—30	25	1.39794	27	37.74438
30—40	35	1.54407	21	32.42547
40—50	45	1.65321	15	24.79815
Total			100	131.80365

Here $\Sigma f = N = 100$, $\Sigma f \log x = 131.80365$

$$\text{G.M.} = \text{Antilog} \left\{\frac{1}{\Sigma f}[\Sigma f \log x]\right\}$$

$$= \text{Antilog}\left[\frac{131.80365}{100}\right] = \text{Antilog}\ [1.318036] = 20.7987$$

Use of Geometric Mean in Computing Rate of Growth

Geometric mean provides a satisfactory measure for computing rate of growth of population phenomena, specially the phenomena which grow at geometric progression

The formula is

$$P_t = P_0\,(1+r)^t$$

or

$$r = \left(\frac{P_t}{P_0} - 1\right)^{\frac{1}{t}}$$

where P_t = the value of variable at the end of the period, i.e., at time t.

P_0 = the value of the variable at the beginning.

r = the rate of growth per unit of time.

t = number of units of time.

Example 1.13

Population of India in 1961 and 1971 were 43.9 and 54.8 crores. Find the rate of increase.

Solution :

Here $P_1 = 54.8$; $P_0 = 43.9$, $t = 10$ years; $r = ?$

$$r = \left(\frac{P_t}{P_0} - 1\right)^{\frac{1}{t}} = \left(\frac{54.8}{43.9} - 1\right)^{\frac{1}{10}}$$

or $\text{V}\log r = \frac{1}{10}\log\left(\frac{54.8}{43.9} - 1\right) = \frac{1}{10}\ (.248292)$

r = Antilog (.0248292) = 1.05884

1.5 Harmonic Mean

In problems such as work, time and rate, where the amount of work is held

constant and an average rate is required, the harmonic mean (H.M.) is utilized. It is defined as the reciprocal of the arithmetic mean of the reciprocals of the given individual readings, i.e., H.M., of $X_1, X_2, \ldots . X_n$ is defined as :

$$H.M. = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots . + \frac{1}{X_n}} \qquad (X_i > 0)$$

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

where n is the number of observations.

Example 1.14

The H.M. of 2, 4, 6 is

$$\frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{6}} = \frac{36}{11} = 3.37$$

In case of frequency distribution (grouped data), if $f_1, f_2, \ldots, f_n$ be the frequencies corresponding to $X_1, X_2, \ldots . X_n$ then H.M. is computed as :

$$H.M. = \frac{f_1 + f_2 + f_3 + \ldots . + f_n}{\frac{f_1}{X_1} + \frac{f_2}{X_2} + \ldots . + \frac{f_n}{X_n}} \quad ; \qquad (X_i > 0)$$

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

Example 1.15

Compute H.M. from the data given in example 9.

Calculation:

Class - group *(X)*	Mid-values *(m)*	No. of observations *(f)*	$\frac{f}{m}$
0—10	5	14	2.80
10—20	15	23	1.53

20—30	25	27	1.08
30—40	35	21	0.60
40—50	45	15	0.33
Total		100	6.34

Here $\Sigma f = N = 100$; $\Sigma(\frac{f}{m})$ 6.34

$$\text{H.M.} = \frac{\Sigma f}{\Sigma(\frac{f}{m})} = \frac{100}{6.34}$$

$= 15.77$

1.6 Median

Median is another important and useful measure of central tendency. It has the connotation of the middle most or most central value of a set of measurements. It is usually defined as the value which divides a distribution in such a manner that the number of items below it is equal to the number of items above it. The median is thus a positional average. It is better indication of central tendency when one or two of the peripheral readings are too large or too small because they give the wrong idea of the average when mean is computed.

Median is that variate value of the data or frequency distribution which divides it in two equal halves.

1.6.1 Calculation of Median (Ungrouped data)

Case 1(n is odd) : In case of ungrouped data when the number of observations are odd, then median is the middle value after the measurements have been arranged in ascending or descending order of magnitude, i.e., if there are n number of measurements and measurements are arranged in ascending or descending order of magnitude, the median of the measurements is $\left(\frac{n+1}{2}\right)^{th}$ measurement where n is an odd number.

Case 2 (n is even) : If the number of observations are even, median is defined as the mean of the two middle observations when observations are arranged in ascending or descending order of magnitude i.e.

$$\text{median} = \frac{\left(\frac{n}{2}\right)^{th} \text{ value} + \left(\frac{n}{2}+1\right)^{th} \text{ value}}{2}$$

Example – 1.16

Calculate median for the following data :

(a) 68, 62, 75, 82, 68, 71, 68, 71, 62, 68, 74, 59, 74, 68, 60, 71, 59, 73, 73, 58.

(b) 200, 150, 260, 285, 380, 305, 4989, 307, 1280, 233, 403,

Solution (a)

To compute the median, first we arrange the values in ascending order of magnitude as :

58, 59, 59, 60, 62, 62, 68, 68, 68, 68, 68, 71, 71, 71, 73, 73, 74, 74, 75, 82.

The number of observations n is even in this case, i.e., n = 20.

So

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} value + \left(\frac{n}{2}+1\right)^{th} value}{2} = \frac{10th\,value + 11th\,value}{2}$$

$$= \frac{68+68}{2} = 68$$

(b)

Let us first arrange the values in ascending order of magnitude as :

150, 200, 233, 260, 285, 305, 307, 380, 403, 1280, 4989.

The number of observations n in this case is odd, i.e., n = 11 so the median is the $\left(\frac{n+1}{2}\right)^{th}$ value i.e., $\left(\frac{11+1}{2}\right)^{th}$ or 6th value of the observation and thus underlined value., i.e., 305 is the median.

1.6.2 Calculation of Median (Grouped Data)

In case of discrete frequency distribution, median can be obtained with the help of cumulative frequencies as follows :

(i) First find N/2 where $N = \Sigma f$.

(ii) Find the cumulative frequency just greater then $N/2$.

(iii) Corresponding value of X (i.e. of variable) is median.

Example 1.17

Calculate the median height from the data given in example 1.4.

Calculation

Height (in metre) (X)	No. of Units (f)	Cumulative frequency (f_c)
200	142	142
600	265	407
1000	560	967
1400	271	1238
1800	89	1327
2200	16	1343
Total	1343	

Here $f = N = 1343$, $\frac{N}{2} = 671.5$

The cumulative frequency just greater than 671.5 is 967 and corresponding to this cumulative frequency, the value of X is 1000 and thus the median height is 1000 metres.

Median (Continuous Grouped Data) : Median for such distribution is computed by the following formula

$$\text{Median (Md)} = l_m + \frac{\left(\frac{N}{2} - f_c\right)}{f_m}.h$$

where l_m is the lower limit f_m is the frequency of the median class, fc is the

cumulative frequency of the class, preceding the median class and h is the class-width of the median class and $N = \sum f_i$

Example 1.18

Calculate median for the following grouped data.

Interval :	35-45	45-55	55-65	65-76	65-85
Frequency :	2	3	5	1	1
Cum Freq.	2	5	10	11	12

The cum. Freq. is computed from the given freq. dist.

The median position = (n+1)/2 = (12 + 1) / 2 = 6.5

Median lies between observation 55 and 65. Both of these observations fall in category 3, i.e. in class (35-65) with cumulative frequency of 10. Therefore,

$$Median = l_m + \frac{\frac{N}{2} - f_c}{f_m} \times h$$

Where $l_m = 55, N = 12, f_c = 5, f_m = 5, h = 10$

$\therefore$ Median = $55 + \frac{[(12/2) - 5]}{5} \times 10 = 55 + 2 = 57$

Example 1.19

The following table gives the size of land holding of families in a village. Find out the median holding size.

Area of land :

(in acres)	5—9	10—14	15—19	20—24	25—29	30—34
No. of families	20	35	150	70	44	38

Solution :

Since the class groups are given in the discrete form hence we first have to convert it into continuous form by adding .5 to the upper limits and substracting .5 from the lower limits as given in column 2 below :

Area of land (in acres) (X)	Area of land (in acres) Continuous from	No. of families (*f*)	Cumulative frequency (*f*)
5—9	4.5 — 9.5	20	20
10—14	9.5 — 14.5	35	55
15 — 19	14.5 — 19.5	150	205
20 — 24	19.5 — 24.5	70	275
25 — 29	24.5 — 29.5	44	319
30 — 34	29.5 — 34.5	38	357
Total		357	

Here $\frac{N}{2} = 178.5$, the cumulative frequency just greater than 178.5 is 275 and the corresponding class group is the median class. For this median class, we have

$l_m = 14.5, f_c = 55, f_c = 150$ and $h = 5$

$$\text{Median} = 14.5 + \left(\frac{178.5 - 55}{150}\right) \times 5$$

$$= 14.5 + 4.12 = 18.62 \text{ acres.}$$

1.6.3 Calculation of Median by (Graphical Method)

One of the methods to compute median is the graphical method. In this case take the class intervals (or the individual readings) on the axis of X and plot the corresponding cumulative frequencies on the axis of Y against the upper limit of the class interval (or against the variate value in case of discrete frequency distribution). The curve obtained by joining the points by means of free hand drawing is the cumulative frequency curve or ogive. For the calculation of median, take a point on the axis of Y that is equivalent to N/2 and from this point draw a line parallel to X-axis. This line will cut the curve and from the cutting point draw a line perpendicular on X-axis. The distance from origin to the point at which the perpendicular line cuts the X-axis is the value of median.

Example 1.20

Find out the median rainfall from the distribution given in example 1.13 by the graphical methods :

Figure 3.1 shows the cumulative frequency curve formed between the upper limits of classes and corresponding cumulative frequencies. The point *N/2* is shown on Y-axis and the dotted line parallel to X-axis cuts the cumulative curve at C. Perpendicular line cuts the X-axis at M. The distance *OM* is the median value.

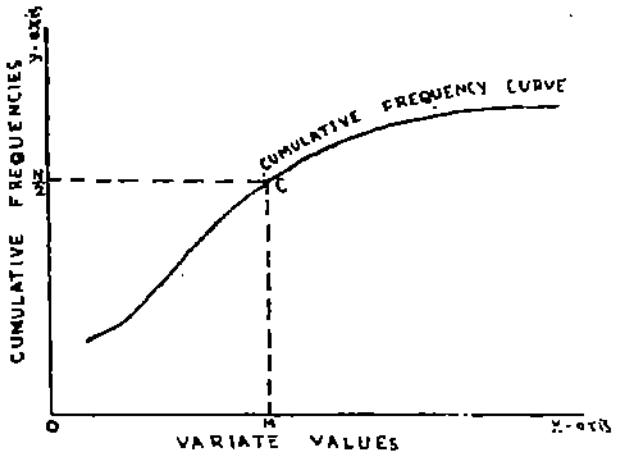


FIG. 3.1 : Cumulative Frequency Curve

1.6.4 Advantages of Median

An important advantage of median is that it is less sensitive than the mean to extreme scores. For skewed data, the median is a better choice because it is usually not affected by a few outlier. Median is also a desirable measure when the distribution has to be turncated for some reasons. If the purpose is to describe the central tendency of a set of scores, the median is preferable to other measures. It gives an undistorted picture of central tendency whether the data are skewed or not.

1.6.5 Disadvantages of Median

Under usual circumstances, the median is more vulnerable to sampling variability than the arithmetic mean. This makes median less stable than the mean from sample to sample and therefore it is not very useful in inferential statistics. For ordinal data, median also ignores the actual values of the observations and simply takes into account their positions.

1.7 Mode

Mode or modal value of a distribution is that value which occurs most frequently. For example, at any station the average number of occurrences for

thunder storms or days with snowfall, wind direction, etc. are the most realistically presented by modal value. In case of frequency distribution the mode is that value which has maximum frequency. If two or more observations occur the same number of times then there is more than one mode and the distribution is called multi-modal as against uni-model.

Types of Measures of Central Tendency

The measures of central tendency or averages are of different types, but the most common in use are of three types:

1. Mean
2. Median
3. Mode

The mean is further classified as :

(i) Arithmetic mean

(ii) Geometric mean

(iii) Harmonic mean

Since each one of the above measures of central tendency has its own individual characteristics and properties, a decision must always be made as to which would be the most appropriate and useful in view of the nature of the statistical data and purpose of the inquiry. The qualities desired in a measure should be (a) rigidly defined, (b) easily computed, (c) capable of a simple interpretation, (d) not unduly influenced by one or two extremely large or small values, and (e) likely to fluctuate relatively little from one random sample to another (of the same size and from the same population). However, the decision about which of the three measures of central tendency to use will be clear after learning the computation of each one. A few general considerations in choosing a measure of central tendency are : (i) the purpose of research – what characteristics of the data are of interest; (ii) the level of measurement of the data- nominal, ordinal, interval, or ratio level; (iii) the shape of the frequency distribution as indicated by a graph – symmetric or skewed; (iv) level of expertise of the researcher and the audience – what can you accomplish and what your audience is able to understand.

1.7.1 Calculation of Mode (ungrouped data)

Mode is defined as that variate value of the data or the frequency distribution which occurs most frequently.

The mode in a series of individual measurements can be located either of two ways.

(i) Data should first be placed in an array so that repetition of a value can be identified and quickly counted, the value of that item which occurs most of the times is the modal value.

(ii) Data should be converted into a discrete series.

Example 1.21

Find the modal temperature value from the values given in example 1.14.

Solution (i) Putting data in array as :

58, 59, 59, 60, 62, 62, 68, 68, 68, 68, 68, 71, 71, 71, 73, 73, 74, 74, 75, 82.

Here mode = 68° F.

(ii) Discrete series (converted to frequency distribution form)

variable (X) :	58,	59	60,	62,	68,	71,	73,	74,	75,	82
Frequency (f):	1	2	1	2	5	3	2	2	1	1

Here the value 68 occurs the maximum number of times, hence it is mode.

1.7.2 Discrete Series (ungrouped data)

In case of discrete frequency distribution, mode can be located by inspection of the distribution alone. The size having the maximum frequency will be reckoned as mode.

Example 1.22

Compute the model size of children born per family in the locality from the data given in example 1.3.

Solution :

The highest size of frequency in the given distribution is 154 and corresponding to this frequency the number of children born per family is 2.

Hence the modal size of children born per family in the locality is 2.

1.7.3 Continuous Series (grouped data)

(i) Determine the modal class interval. It is the class interval with the maximum number of frequencies in it. This can be found out by just observing the series.

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

$$\text{Mode} = L + \left(\frac{f - f_p}{2f - f_p - f_s} \right) h$$

where

L is the Lower limit of the modal class; f is the frequency of the modal class; f_p is the frequency of the class preceding the modal class; f_s is the frequency of the class succeeding the modal class and h is the class width of the modal class.

Example 1.23

Compute the modal agricultural holding of the village from the data given in example 1.19,

Solution:

The maximum frequency 'f' in the distribution is 150 which corresponds to class group 15-19, i.e. 14.5—19.5 in continuous case (see column 2, example 17). Hence modal class is 14.5—19.5. Now mode is computed as :

$$\text{Mode} = L + \left(\frac{f - f_p}{2f - f_p - f_s} \right) \times h$$

Here $L = 14.5, f = 150$, $f_p = 35$, $f_s = 70$ and $h = 5$.

$$\text{Mode} = 14.5 + \frac{150 - 35}{150 \times 2 - 35 - 70} \times 5$$

$$= 14.5 + \frac{115}{195} \times 5 = 17.45 \text{ acres.}$$

Sometimes mode is also computed with the help of mean and median. For a symmetrical distribution mean, median and mode coincide

and if the distribution is moderately asymmetrical, the mean, median and mode are approximately related by the formula :

Mode ≅ 3 Median — 2 Mean.

Example 1.24

If the mean and median of a moderately asymmetrical series are 12.9 and 12.1, respectively, what would be its most probable mode ?

Solution:

Mean = 12.9, Median = 12.1, Mode ?

Mode = 3 Median — 2 Mean

= 3 x 12.1 — 2 x 12.9 = 36.3 — 25.8 = 10.5

1.8 Percentiles, Deciles And Quartiles

In general, the term 'fractile' refers to a score that divides a distribution into fractional parts. Percentiles is the most commonly used fractile. Other fractiles are deciles, quartiles, etc. Percentiles, deciles, and quartiles are used as measures of location—to locate the position of a score relative to other scores in a distribution. These measures are close relatives to median.

Percentile is a score below which a certain percentage of scores falls. For example, a student falling at the ninety-first percentile on an examination means that 91 percent of the students had scores lower than his/her.

Percentile score is the raw corresponding to a percentile rank, where percentile rank is the rank (from 0 to 100) at which a particular raw score falls. A percentile rank tells us the percentage of scores falling below a score. Such a score is referred to as percentile.

Percentiles divide the distribution into 100 portions of equal size. For example, sixty-fifth percentile is the score below which 65% of the cases fall. Similarly, deciles divide the distribution into 10 portions of equal size. For example, third decile is the score below which 30% of the scores fall. Calculating the third decile is equivalent to calculating the 30th percentile. Quartiles divide the distribution into 4 portions of equal size, For example, second quartile is the point or score below which 50% of the scores fall. Calculating the 2nd quartile is equivalent to the 50th percentile.

Each percentile equals 1% of a distribution : $1 \times 100 = 100$

Each decile equals 10% of a distribution : $10 \times 10 = 100$

Each quartile equal 25% of a distribution : $25 \times 4 = 100$

Notice that the percentile ranks and percentile scores can also be read directly from cumulative frequency graphs or the cumulative frequency column of a frequency table.

1.8.1 Percentile Score from a Given Percentile Rank

Percentiles can be computed using mathematical formula. Make a frequency distribution and locate the interval in which the percentile of interest belongs. This can be done by using the column of cumulative percentage frequencies. Suppose one is interested in finding the value of the pth percentile. Referring to the column of cumulative percentage frequencies, locate the class interval that contains the pth percentile. Then use the following formula to determine the approximate value of the pth percentile.

$$P_p = l_p + \frac{pn/100 - F_p}{f_p} \times w_p$$

where F_p = cumulative frequency upto, but not including, the pth percentile category.

f_p = number of cases in the interval containing pth percentile.

l_p = lower limit of interval containing pth percentile

w_p = width of interval containing pth percentile.

p = percentile rank

n = sample size

To Calculate Percentile Rank Given a Percentile Score

$$p \times \frac{F_x + \left[(X - l_x)/w_x\right] f_x}{n} \times 100$$

where F_x = cumulative frequency up to but not including, the interval containing the X;

X = given raw scores

w_x = width of interval containing X

lx = lower limit of interval containing X

fx = frequency in interval containing X

n = total number of scores or cases.

Example 1.25

Consider the following distribution of number of priseners arrest for 50 inmates at a state prison.

Interval	True limits	*fi*	*cfi*
0—2	–.5–2.5	0	0
3—5	2.5–5.5	17	17
6—8	5.5.–8.5	15	32
9–11	8.5–11.5	8	40
12-–14	11.5–14.5	4	44
15—17	14.5–17.5	3	47
18—20	17.5–20.5	1	48
21—23	20.5–23.5	1	49
24—26	23.5–26.5	0	49
27–29	26.5–29.5	1	50
Total		50	

(i) Find the percentile rank of a prisõner who has been arrested 6 times.

(ii) How many times a prisoner has been arrested in order to be at the (a) 2nd quartile (b) 3rd quartile ?

Solution :

(i) Percentile rank, p_x, from a given score X where X = 6.

$$p_x = \frac{F_x + \left[(X - l_x)/w_x\right] f_x}{n} \times 100$$

where X = 6 (it fall in interval 5.5–8.5); $F_6 = 17$; $w_6 = 3$; $l_6 = 5.5$;

$f_6 = 15$; $n = 50$

$$p_x = \frac{17 + [(6 - 5.5)/3]15}{50} \times 100 = [(17 + 2.5)/50]100 = 39$$

A percentile rank of 39 means 39% of the prisoners were arrested 6 times or less.

(ii) (a) Percentile score X_p for a Given Percentile Rank p of 50.

$$X_p = l_p + \frac{pn/100 - F_p}{f_p} \times w_p$$

where p = 50 (2nd quartile is equivalent to 50th percentile). Also notice that the 50th percentile falls in the interval 5.5–5.8 since 50% of the score are in or below this interval.

$$F_{50} = 17; f_{50} = 15; l_{50} = 5.5; w_{50} = 3$$

$$X_{50} = 5.5 + \frac{(50 \times 50)/100 - 17}{15} \times 3 = 5.5 + (8/15)(3)$$

$$= 5.5 + 1.6 = 7.1$$

A percentile of 7.1 means a prisoner should have been arrested about 7 times to be at the 2nd quartile or the 50th percentile.

(ii) (b) Percentile score X_p for a Given Percentile Rank p of 75 :

$$X_p = l_p + \frac{pn/100 - F_p}{f_p} \times w_p$$

where p = 75 (3rd quartile is equivalent to 75th percentile). Also notice that the 75th percentile falls in the interval 8.5-11.5 since 75% of the scores are in or below this interval.

$$F_{75} = 32; f_{75} = 8; l_{75} = 8.5; w_{75} = 3; n = 50$$

$$X_{75} = 8.5 + \frac{(75 \times 50)/100 - 32}{8} \times 3 = 8.5 + (5.5/8)(3) = 10.56$$

A percentile score of 10.56 means a prisoner should have been arrested about 11 times in order to be at the 3rd quartile or the 75th percentile. In other words, 75% of the prisoners were arrested less than 11 times.

1.9 Choosing a Measure of Average

Following are a few important criteria in choosing a measure of average :

(1) If there is a specific purpose or goal in mind, choose a measure of central

tendency that will help to achieve that goal. The measure chosen may or may not be appropriate. An inappropriate measure (calculated in violation of its assumptions) may serve the purpose better than an appropriate measure can.

(2) If the variable is nominal, only mode can be calculated, and thus the choice is simple. If the variable is ordinal, both mode and median can be calculated. But median is a better measure because it makes use of more information. If the variable is interval or ratio level, all three measures can be calculated. In that case, if the distribution is normal (symmetric and bell-shaped), all three measures will have the same value. But if the distribution is skewed, median is a better measure of central tendency because it ignores the extreme scores responsible for causing the skewness.

Relationship Among Measures of Average

For symmetrically shaped distribution : Mean = Median = Mode

For positively-skewed distribution : Mean > Median > Mode

For negatively-skewed distribution : Mean < Median < Mode

1.10 Exercises

1.1. Data are collected on the weekly expenditures of a sample of urban households on food. The data, obtained from diaries kept by each household, are grouped by number of members of the household. The expenditures were as follows:

1 member: 67 62 168 128 131 118 80 53 99 68 76 55 84 77 70 140 84 65 67 183

2 member: 129 116 122 70 141 102 120 75 114 81 106 95 94 98 85 81 67 69 119 105 94 94 92

3 member: 79 82 99 142 171 82 145 94 86 85 100 191 116 100 125 116.

4 member: 139 111 251 106 93 99 155 132 158 62 114 129 108 91.

5 or more members: 121128 129140 206 111 104 109 135 136.

For each number of members calculate the mean, median, mode, 25th percentile,

50th percentile, and 75th percentile. Interpret each statistics.

1.2. An instructor gives a quiz with three questions, each worth 1 point; 40% of the class scored 3 points, 30% scored 2 points, 20% scored 1, and 10% scored 0:

Score	Percent
3	40
2	30
1	20
0	10

(i) If there were ten people in the class, what would the average score be?

(ii) If there were twenty people in the class, what would the average score be?

(iii) Suppose you are not told the number of people in the class. Can you still figure out the average score? Explain.

1.3. *A:* In 1989, Governor Brown of California proposed that all state employees be given a flat raise of $ 70 a month. What would this do to the average monthly salary of state employees? What would a 5% increase in the salaries, across the board, do to the average monthly salary? What will the doubling of salaries do to the mean salary?

B: A politician charges the opposition political party with spending an average of over Rs. 100,000 for its candidates from the state and that this is an outrageo sum for a party to spend on its candidates, specially on candidates for state senator and state representative. The campaign spending figures for the party are:

Office	No. of Candidates	Average Amount Spent (Rs)
U.S. Senator	2	1,000,000
U.S. Congressman	16	400,000
Governor	1	800,000
State Senator	50	35,000
State rep.	50	23,000

Calculate mean, median and mode for campaign spending. Is the politician's

criticism valid ? Explain

1.4. In a corporation, a very small group of employees has extremely high salaries, while the majority of employees receive much lower salaries. If you were the bargaining agent for the employees, what measure of average would you calculate to illustrate the low pay level, and why " If you were the employer, what kind of average would you use to demonstrate a high pay level, and why ?

5. Which measures of central tendency are appropriate for each of the following variables ? If several can be calculated, indicate which makes most use of the available information. Comment briefly on each.

(a) Number of siblings (b) Political party affiliation

(c) Satisfactory with family (d) Vacation days per year

(e) Type of car driven

1.11 Summary

Various measures of central tendency have been defined in this unit. There is found a tendency in the data to cluster around a central value. This value is known as measure of central tendency. These are mean, median and mode. Mean is obtained by dividing the sum of observations by number of observations. Median is that variate value which divides the given data or frequency distribution in two equal halves. Mode is that variate value which occurs most frequently i.e. for which the frequency is maximum. Mean, median and mode approximately satisfy the relation.

Mean – mode = 3 (mean – median)

1.12 Further Readings

1. Goon, A.M., Gupta M.K. and B. Dasgupta : Fundamentals of Statistics, Volume I, the World Press Pvt. Limited, Calcutta.

2. Yule G.U. and Kendall M.G. : An Introduction to the theory of Statistics, Charles Griffin and Co. Limited.

Unit-2 Measures of Dispersion

Structure

2.1 Introduction

The objective of numerical description is to obtain a set of measures that will create a mental reconstruction of the frequency distribution of the data. Now we know something about averages, let us turn to another type of measures, called the measures of dispersion. These measures summarize how spread out scores are. It is useful to know how similar or dissimilar scores are from the average score and from one another. We might like to know if scores cluster, they are more homogeneous. If scores are spread out widely, then they are more heterogeneous. The measures of central tendency as discussed above only locate the center of a distribution, but tell nothing about the degree of variability. In order to study the dispersion or variability in a distribution, we need alternative measures called the measures of dispersion.

Measuring dispersion is important for two reasons. First, quantifying the dispersion in the data is required by many of the statistical inference tests that will be discussed later. Second, dispersion of a distribution, in conjunction with the

central tendency, more completely describes the distribution. Whereas measures of central tendency are quantification of the average value of the distribution, measures of dispersion are quantification of the extent of dispersion.

This can be illustrated with the following sample:

Suppose 20 students, 10 in a sociology class and 10 in a statistics class, are asked how many hours of TV they watched last week. Their answers are as follows:

Sociology (hours) :	4	4	5	5	5	5	5	5	6	6
Statistics (hours) :	0	0	0	0	0	3	10	10	12	15

Mean number of hours for sociology students = (4 + 4 + 5 +... + 6 + 6) /10 = 5.

Mean number of hours for statistics students = (0 + 0 + ... + 12 + 15)/10 = 5.

The average number of hours for the two groups of students is same—5 hours per week. However, there is a difference in the way values are distributed in the two distributions. All of the sociology students watched TV between 4 and 6 hours during the week and there is little variation in the hours from student to student. The statistics students, however, differ from each other much more. Some seem to have devoted themselves entirely away from TV, while others watched a lot of TV. Thus several distributions may have the same mean, but differ from each other in the way scores are distributed.

If central tendency is thought of as the point that best represents a central score in a distribution, the dispersion presents the other side of the coin. Dispersion reflects the "goodness" or "poorness" of central tendency as a representation of all the scores in a distribution.

2.2 Objectives

After studying this unit you will be able to understand :

methods of computing various measures of dispersion;

the advantages as well as limitations of each of these measures;

- the relationship between measures of central tendency and measures of dispersion;
- the coefficient of variation as a measure for comparing two distributions.

2.3 Types of Measures of Dispersion

Dispersion is defined as the degree to which scores deviate from the central tendency (usually the mean) of the distribution. The statistical techniques that quantify this dispersion in a distribution are called measures of dispersion. Most commonly used measures of dispersion are range, average deviation, variance, and standard deviation.

2.4 Range

Range is defined as the difference between the highest and the lowest scores in a distribution. Symbolically,

$$R = X_{max} - X_{min}$$

where R is the range, X_{max} is the highest score, X_{min} is the lowest score.

A large value of range indicates greater dispersion and a small value of range indicates lesser dispersion among the scores. Minimum value that range can achieve is 0 and the maximum is infinity. If all the scores are the same, R will have a value of 0 and hence there is no dispersion.

Example 2

Find range for values : 87, 92, 47, 58, 87, 62, 73, 73, 61.

Solution :

It is always a good idea to first rank the observation in ascending or descending order. In an ascending order, the scores are : 47, 58, 61, 62, 73, 73, 87, 87, 92. A visual examination shows that $X_{max} = 92$; $X_{min} = 47$.

Therefore $R = 92 - 47 = 45$

2.4.1 Advantages of Range

(i) Range gives a quick indication of dispersion. It can be a good measure if there are no outliers in the data that means the distribution is not skewed.

(ii) Range is easy to compute and interpret. For variables measured at an ordinal scale, range is the only measure which is technically meaningful.

(iii) If the data are to be presented to a relatively unsophisticated audience, the range may be the only measure of dispersion that will be readily understood.

2.4.2 Disadvantages of Range

(i) Calculation of range is based only on two extreme scores, the minimum and the maximum. The rest of the data are ignored.

(ii) Range tells nothing about the dispersion among intermediate scores.

(iii) Range is greatly affected by outliers. Thus for skewed distributions, range is usually very misleading measure.

(iv) Since range ignores all the scores except the two extreme scores, it cannot be used for making inferences about populations.

(v) Range varies considerably from sample to sample.

2.4.3 Interquartile Range

The interquartile range, usually denoted by IQR, is a kind of range. It avoids some of the problems associated with R by taking into consideration only the middle half of a distribution. To find IQR :

(i) Arrange the scores from lowest to highest.

(ii) Divide the distribution into quartiles and calculate the first, the second, and the third quartiles using the formulas discussed in the previous unit.

(iii) The IQR is defined as the distance between the third quartile Q_3 and the first quartile Q_1. Symbolically,

$$IQR = Q_3 - Q_1$$

Thus, IQR extracts the middle half of the cases and then calculates the range. IQR avoids the problem of being based on the most extreme scores by excluding the two extremes, but it has all the other disadvantages associated with

R. For example, Q fails to yield any information about the nature of scores other than Q_3 and

2.5 Mean Deviation

Average deviation is found by summing the absolute values of the deviations and dividing the sum by number of observations. The formula for average deviation can be written as :

$$MD = \left(\Sigma| X_i - \overline{X} |\right) / n$$

where $\overline{X}$ is the arithmetic mean, $X_i - \overline{X}$ is deviation of $\overline{X}_i$ from $\overline{X}$, and $|X_i - \overline{X}|$ is the absolute value of the deviation which is always a positive number. The average deviation tells the distance with which a score will typically deviate from the mean.

Example 2.2

The number of terms that five randomly selected Members of Parliament have served are : 3, 10, 12, 7, 8. Find the average deviation of these scores.

Solution :

Make the following table containing the calculations.

Case number	Terms	$X_i - \overline{X}$	$\lvert X_i - \overline{X} \rvert$
1	3	$3 - 8 = -5$	5
2	10	$10 - 8 = 2$	2
3	12	$12 - 8 = 4$	4
4	7	$7 - 8 = -1$	1
5	8	$8 - 8 = 0$	0
Total	40	0	12

Mean $\overline{X} = 40/5 = 8$

Sum of deviation from the mean $(X_i - \overline{X}) = 0$

Sum of absolute deviations $|X_i - \overline{X}| = 12$. Therefore

$$MD = 12 / 5 = 2.4 \text{ terms.}$$

For descriptive purposes, the average deviation can be an adequate and easily interpretable measure for describing the degree of dispersion. But the mathematical properties of average deviation are such that it does not meet the needs of advanced mathematics. Therefore, average deviation is a very infrequently used measure of dispersion.

2.6 Variance and Standard Deviation

Instead of taking the absolute values of deviations to remove the negative signs and obtain a nonzero sum, another way to get rid of negative sign of deviations is to square them. Square of a negative number is a positive quantity. A statistic called variance uses this approach. To calculate variance, calculate the deviations, square each deviation, add up the squared deviations to obtain sum of squares, and divide this sum of squares by the number of deviations. The resulting quantity is called the mean squared deviation (MSD) or the variance of the distribution of scores.

Variance can be of two types:

(i) Sample variance, denoted by S^2, calculated from sample data.

(ii) Population variance, denoted by σ^2 calculated from population data.

In practice S^2, a statistics, is always known while σ^2, a parameter, is seldom known. Therefore, S^2 is used as an estimate of σ^2. While dealing with several variables, it proves to be convenient to attach a subscript to s or σ. The subscript indicates the name of variable for which variance is being calculated Thus, s_x^2 is the sample variance of the variable X, s_y^2 is the sample variance of Y, and so on.

2.6.1 Computation of Variance

Variance may be defined as the mean squared deviation of scores around the mean. In the form of a formula, variance is given by:

$$s_x^2 = \Sigma \left(X_i - \overline{X}_i\right)^2 / n \quad \text{(for sample data)}$$

$$\sigma_x^2 = \Sigma \left(X_i - \mu\right)^2 / N \quad \text{(for population data)}$$

where s_x^2 = sample variance of variable X; σ_x^2 is population variance; X_i is the value of X variable for ith case; $\overline{X}$ is sample mean; μ is population mean; n is the sample size; and N is population size.

The above formula is used only when each score has a frequency of 1 and data are ungrouped. If some scores occur more or less frequently than others, a compact ungrouped frequency table may be constructed in which the entries in the column of frequencies are not all the same and they are not all 1's. In such a case, the formula for variance is written as:

$s_x^2 = \Sigma f_i (X_i - X)^2 / (n-1)$ where f_i is the frequency of X_i

$\sigma_x^2 = \Sigma f_i (X_i - \mu)^2 / N$ where f_i is the frequency of X_i

In calculating the sample variance, the reason for dividing by n-1, instead of n, is to get an unbiased estimate of known population variance. The variability of a sample of scores tends to be less than the variability of the population from which the scores are taken. In order to use the sample variance as an unbiased estimate of population variance, a correction factor (n-1) is used in the denominator of the formula for the variance of a sample. In other words, sample variance almost always underestimates its corresponding population variance and dividing by (n-1), instead of n, tries to compensate for this underestimation. For larger sample sizes (n > 100), it makes little difference whether one divides by n or n-1. Significant error can occur if sample is small (n < 25). In situations where the interest is merely in describing the variability in the data at hand, only n should be used as a divisor. As a general rule, one can almost always use n-1 for sample data.

As a descriptive statistic for variability, the variance changes in value as a function of the amount of variability in the data. When all scores are identical, the value of variance will be zero. As scores become more dispersed around the mean, the value of variance increases. Variance is based on squared deviations and, therefore, it is always greater than or equal to zero.

2.6.2 Standard Deviation

Although variance is a very useful measure of variability, its value as a descriptive statistic is limited somewhat by the difficulty most people have in thinking about squared deviations. For instance, if you were calculating the variability for income scores (measured in Rs.), the variance will be expressed in

squared units (Rs. Rs. or Rs.^{2}) and you might obtain a value of variance, say 16 Rs. Rs.. In the process of squaring the deviations, the units also get squared. This is, what is done, when area is reported and calculated—square feet, square inches, etc.

Computing the square root of the variance expresses this variability in terms of the original score values, such as Rs. 4, which is easier to interpet and comprehend. This square root of the variance is called the standard deviation (SD), represented by s. The SD is approximately equal to the mean deviation (MD) of scores around the mean. Since variance is the mean squared deviation (MSD), standard deviation is root mean squared deviation (RMSD). Because the standard deviation is more readily interpretable than variance, it is used more often to describe data variability.

That is

Standard deviation = Square root of variance or $s = \sqrt{s^2}$

$$s = \sqrt{s^2} = \sqrt{\left[\Sigma \left(X_i - \overline{X}\right)^2 / (n-1)\right]} = \sqrt{\left[\Sigma f_i \left(X_i - \overline{X}\right)^2 / (n-1)\right]}$$

Similarly, population standard deviation $\sigma = \sqrt{\sigma^2}$

2.6.3 Effect of Change of Origin and Scale

The S.D. is.

$$S_x = \sqrt{\frac{\Sigma (X_i - \overline{X})^2}{(n-1)}}$$

where $\overline{X} = \frac{1}{n} \Sigma x_i$

If $\delta_i = \frac{x_i - a}{h} \quad \forall i = 1,2,\ldots n.$

$$\Rightarrow x_i = \delta_i h + a$$

Then $\overline{X}_{\delta_i} = a + \frac{h \Sigma \delta_i}{n}$

and $(n-1)S^2 = \sum \left[(\delta_i h + a) - \left(a + \frac{h \Sigma \delta_i}{n} \right) \right]^2$

$$= \sum \left[\delta_i h - h \left(\frac{\Sigma \delta_i}{n} \right) \right]^2$$

$$= h^2 \sum (\delta_i - \bar{X}\delta_i)^2$$

$$S_{x_i}^2 = h^2 \frac{\sum (\delta_i - \bar{X}\delta_i)^2}{n-1}$$

$$= h^2 S_{\delta_i}^2$$

$$S_{xi} = \sqrt{h^2 S_{\delta i}^2} = h S_{\delta i}$$

$$S_{xi} = h S_{\delta i}$$

i.e. standard deviation is independent of change of origin but not independent of change of scale.

Example 2.3

Calculate the mean and S.D. for the following given table of marks distribution of 50 students.

Marks	Students (f)	Mid value (x)	$x_i - 25$
0-10	2	5	5-25 = -20
10-20	10	15	15-25 = -10
20-30	15	25	25-25 = 0
30-40	14	35	35 - 25 = 10
40-50	9	45	45 - 25 = 20
	50		

$\delta_i = \frac{x_i - 25}{10}$	$f_i \delta_i$	$f_i \delta_i^2$
-2	-4	8
-1	-10	10
0	0	0
1	14	14
2	18	36
	18	68

$$\bar{x}\,\delta_i = a + h\frac{\Sigma f_i \delta_i}{N} \qquad N = \sum f_i$$

$$= 25 + \frac{10 \times 18}{50} = 25 + 3.6 = 28.6$$

$$S_{x_i}^2 = h^2 S^2$$

$$S_\delta^2 = \left(\frac{1}{N}\sum f_i \delta_i^2\right) - \left(\bar{x}_\delta^2\right)$$

$$= \left(\frac{1}{N}\sum f_i \delta_i^2\right) - \left(\frac{1}{N}\sum f_i \delta_i\right)^2$$

$$= \frac{1}{50} \times 68 - \left(\frac{1}{50} \times 18\right)^2$$

$$= 1.36 - (0.36)^2$$

$$= 1.36 - 0.1296 \qquad = 1.4896$$

$$S_{x_i}^2 = 10 \times 1.49 = 14.896$$

Example 2.4

For a group of 100 candidates, the mean and standard deviation were found to be 20 and 8 respectively. Latter it discovered that No. 23 and 36 were missread as 32 and 63. Find the correct mean and S.D. Corresponding to the correct numbers.

Solution :

Let x be the variable . we have

$n = 100 \qquad \bar{x} = 20 \qquad S = 8$

Now

$$\bar{x} = \frac{1}{n}\Sigma x_i \qquad \Rightarrow n\bar{x} = \Sigma x_i$$

$$\Rightarrow \Sigma x_i = 100 \times 20 = 2000$$

$$corrected\ \Sigma x_i = 2000 - 23 - 36 + 32 + 63 = 2036$$

$$Correct\ mean = \frac{2036}{100} = 20.36$$

Similarly,

$$S^2 = \frac{1}{n}\Sigma x_i + \bar{x}^2$$

$$\Sigma x_i^2 = n(S^2 + \bar{x}^2)$$

$$\Sigma x_i^2 = 100(64 + 400) = 46400$$

Corrected

$$\Sigma x_i^2 = 46400 - (23)^2 - (36)^2 + (32)^2 + (63)^2$$

$$= 46400 - 529 - 1296 + 1024 + 3969$$

$$= 49568$$

Corrected $\Sigma x_i^2 = 49568.$

Now corrected $S^2 = \frac{49568}{100} - (20.36)^2$

$= 495.68 - 414.5296 \quad = 81.1504.$

Corrected $S = 9.0084 \quad \cong 9$

Corrected mean = 20.36

Corrected S = 9.

2.6.4 Steps in Computing the Standard Deviation

(i) Make a frequency distribution table, if not already made, containing two columns, namely, the columns for score values and their frequencies.

(ii) Calculate the mean score, if not already given :

$$\overline{X} = \Sigma f_i X_i / n \quad \text{where } n = \Sigma f_i$$

(iii) Subtract the mean $\overline{X}$ from each of the scores X_i to calculate deviations $X_i - \overline{X}$. Write these deviations in a separate column, say column 3. Sum all these deviations and see if the sum is zero (excepting the rounding errors.)

(iv) Square each deviation obtained in step (iii) and write the squared amounts in a separate column, say column 4.

(v) Sum all entries in col-4 to obtain a quantity $\Sigma f_i (X_i - \overline{X})^2$.

(vi) Take the square root of variance in step (v) to obtain the standard deviation.

Example 2.5

"How accurate are eyewitness reports of accidents ?" Social scientists have studied this question in detail. In one experiment, subject viewed a film of an accident in which a car ran a stop sign and hit a parked car. The speed of the car was 31 miles per hour. After viewing the film, subjects were asked to estimate the speed of the car. Ten subject gave the following estimates :

15, 40, 32, 18, 35, 20, 37, 35, 28, 40

Calculate the mean and standard deviation for these data. How accurate were the estimates considering the mean score across all subjects ? How does the SD help to interpret the mean ?

Solution :

To calculate SD. it is useful to make the following table

Xi	Fi	$fiXi$	X_i^2	$f_iX_i^2$
15	1	15	225	225
40	2	80	1600	3200
32	1	32	1024	1024
18	1	18	324	324
35	2	70	1225	2500
20	1	20	400	400
37	1	37	1369	1369
28	1	28	784	784
Total	10	300		9826

Mean $\overline{X} = \Sigma f_i X_i / n = 300 / 10 = 30$

Sample variance

$$s_x^2 = \left[n\left(\Sigma f_i X_i^2\right) - \left(\Sigma f_i X_i\right)^2\right] / (n)(n-1)$$

$$= [10(9826) - (300)^2] / (10)(10\text{-}1) = (98260 - 90000) / (10)(9)$$

$$= 8260 / 90 = 91.78$$

Standard deviations = $\sqrt{s^2} = \sqrt{(91.78)} = 9.58$

Both variance and standard deviation are based on two important properties of the mean : (i) Sum of the differences of scores from the mean in a distribution equals zero. It is due to this property that the deviations from the mean are squared. (ii) The sum of the squared differences of each value in a distribution from the mean of the distribution yields a minimum value , $\Sigma f_i \left(X_i - \bar{X}\right)^2$ = minimum.

2.6.5 Combined Variance

If n_1 and n_2 be the sizes of two series with respective means $\bar{x}_1, \bar{x}_2$ and respective variances S_1^2, S_2^2, then the standard deviation of the combined series is denoted as S and defined as

$$S^2 = \frac{1}{n_1 + n_2}\left[n_1\left(S_1^2 + d_1^2\right) + n_2\left(S_2^2 + d_2\right)\right]$$

where

$$d_1 = \bar{x}_1 - \bar{x}, \quad \bar{x}_1 = \frac{1}{n_1}\Sigma x_{1i}$$

$$d_2 = x_2 - \bar{x}, \quad \bar{x}_2 = \frac{1}{n_2}\Sigma x_{2i}$$

$$and\ \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}\ (Combined\ mean)$$

another formula is -

$$S^2 = \frac{1}{n_1 + n_2}\left[n_1 S_1^2 + n_2 S_2^2 + \left\{\frac{n_1 n_2}{n_1 + n_2}\left(\bar{x}_1 - \bar{x}_2\right)^2\right\}\right]$$

where

$$S_1^2 = \frac{1}{n_1}\Sigma(x_{1i} - \bar{x}_1)^2$$

$$S_2^2 = \frac{1}{n_2}\Sigma(x_{2i} - \bar{x}_2)^2$$

Example 2.6

An analysis of monthly wages paid to the engineers in two companies A and B gives the following results.

	A	B
No. of engineers	1000	2000
Average monthly salary	240.00	275.00
Variance of distn. of salary	41	80

(a) Calculate average monthly salary.

(b) Variance of the distribution of monthly salary of all engineers in A & B taken together.

Solution:

$n_1 = 1000 \qquad n_2 = 2000$

$\bar{x}_1 = 240 \qquad \bar{x}_2 = 275$

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$= \frac{(240 \times 1000) + (2000 \times 275)}{1000 + 2000}$$

$$= \frac{240000 + 550000}{30000}$$

$$= \frac{790000}{3000} = \frac{790}{3} = 263.34$$

$$= 263.34$$

$$(b) \quad S^2 = \frac{1}{n_1 + n_2}\left[n_1S_1^2 + n_2S_2^2 + \left\{\frac{n_1n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^2\right\}\right]$$

$$= \frac{1}{3000}\left[41000 + 160000 + \left\{\frac{4000000}{3000}(240 - 275)^2\right\}\right]$$

$$= \frac{1}{3}\left[201 + \left\{\frac{4}{3} \times 1225\right\}\right]$$

$$= \frac{1}{3}[201 + 1633.34]$$

$$= \frac{1834.34}{3} = 611.45$$

$S = 24.73$

2.6.6 Properties of Standard Deviation

(i) Standard deviation gives a measure of dispersion relative to the mean.

(ii) Standard deviation is sensitive to each of the scores in the distribution.

(iii) Like the mean, standard deviation is stable with regard to sampling fluctuations. This property is one of the main reasons why the standard deviation is used so much more often than other measures of dispersion.

2.6.7 Interpretation of Standard Deviation

Standard deviation measures the average dispersion in a data set. It is the average amount by which scores in a distribution deviate from the mean of the distribution. Intuitively, large values of standard deviation show that the observations are quite spread out and smaller values indicate that the scores are less dispersed and are clustered around the mean. In an extreme case, for example, a standard deviation of 0 means all of the scores are exactly equal to the mean score and there is no variability among the scores in a distribution.

2.7 Relationship between Measures of Central Tendency and Measures of Dispersion

If $\overline{X}$ be the arithmetic mean, G be the geometric mean, H be the hormonic mean, S^2 be the variance (or S be the standard deviation) and Mol be the mean deviation then for the discrete distribution:

(*i*) $G = \bar{x}\left(1 - \frac{1}{2}\cdot\frac{S^2}{\bar{x}^2}\right)$

(*ii*) $\bar{x}^2 - G^2 = S^2$

(*iii*) $H = \bar{x}\left(1 - \frac{S^2}{\bar{x}^2}\cdot\right)$

(*iv*) $S^2 \geq (Ma\ from\ mean)^2$

2.8 Coefficient of Variation (CV)

It is sometimes desirable to compare several groups with respect to their relative homogeneity in instances where the groups have very different means.

Therefore it might be somewhat misleading to compare the absolute magnitudes of the standard deviations. One might expect that with a very large mean one would find a fairly large standard deviation. One might, therefore, be primarily interested in the size of the standard deviation relative to that of the mean. This suggests that we can obtain a measure of the relative variability by dividing the standard deviation by the mean. The result has been termed the coefficient of variation, denoted by CV. Thus

$$CV = S/\overline{X}$$

where s is the SD and $\overline{X}$ is the mean.

The coefficient of variation, being a ratio, requires that one have a ratio level of measurement and not merely interval measurement. You can also realize that one should always report the mean as well as the SD for the data.

To illustrate the advantages of CV over the SD, suppose a social psychologist is attempting to show that for all practical purposes two groups are equally homogeneous with respect to age. In one group the mean age is 26 with an SD of 3. In the other one, the mean age is 38 with an SD of 5. The coefficients of variation for the two groups are:

$$CV1 = 3/26 = .115,\ CV2 = 5/38 = .132$$

The difference between the two coefficients is smaller than the difference between the two SDs. In view of the fact that exact age usually becomes less important in determining interests, abilities, and social status as the average age of group members is increased, a comparison of the two coefficients of variation in this instance might very well be much less misleading than if the SDs were used.

As another example, suppose one is concerned about the dispersions in traffic flows from one weekday to the next, at various times of the day. Dispersions in these flows might be misleading in an absolute sense unless standardized by their means so as to allow for differences in the average volumes of traffic at different times of the day.

2.9 Exercises

1 You have just won the state lottery and are now fabulously wealthy. One of the first things you want to do is to find the "nicest place to live" in all the world. Because you are somewhat eccentric, your only criterion for "nicest place" is climate. Specifically, you want to locate a city where the

temperature is exactly 78°. After much search, you find three cities where the average daily temperature is exactly 78°. Based on just this much information. which of the cities will you choose as your permanent residence? Now suppose you also discovered that the SO and range of the daily temperature were 0.7° and 3° in city A. 10.3° and 30" in city B, 25.8" and 103° in city C. How will this additional information be useful to you in choosing the place you want to live? Can you choose a permanent residence now? Which city would you choose and why?

2 Suppose, men and women have about the same distribution of scores on the verbal scholastic aptitude test (SAT), but, on the mathematical part. men have a distinct edge. In 1994, the men averaged about 500 on the mathematical SAT. while the women averaged about 460. Both histograms follow the normal curve, with standard deviation of 100.

(a) Estimate the percentage of men getting over 600 on this test in 1993.

(b) Estimate the percentage of women getting over 600 on this test in 1993.

(c) Suppose one of the men who took the mathematical SAT will be picked at random, and you have to guess his test score. You will be given a dollar if you guess it right to within 50 points. What should you guess? What is your chance of winning? Briefly explain your answer.

3 For a normal curve. answer the following questions:

(a) Find the proportion of the area between the mean and a z score of .5.

(b) What proportion of the area lies to the right of a z score of .5?

(c) What proportion of the area lies to the left of a z score of -.5? .

(d) What proportion of area lies in interval bounded by z scores of -1 and -2?

(e) Assume the age of state governors in the United States is normally distributed with a mean of 56 and a standard deviation of 8 years. How many governors are between 60 and 70 years of age? (there are 50 governors total).

4 Suppose that, for a particular year. on the law school admissions test (LSAT). the mean score for all people taking the test is 500. the standard deviation is 90, and the scores are normally distributed. (a) What percentage

of people had scores (i) over 600; (ii) less than 300; (iii) between 700 and 750? (b) If a person had a score of 630 on the test, what percentage of people had scores less than his? greater than his? (c) If 5000 people took the test that year, how many had a score Between 300 and 400?

P-1.6 Suppose that a college entrance examination is given to all entering college students. It is found that the scores are normally distributed with a mean of 450 and a standard deviation of 75. (a) What is the probability that if a student were selected at random, his score in the test would be (i) greater than 450; (ii) less than 550; (iii) between 350 and 400? (b) If the z score of a student on this test were -1.5, what was his original score?

2.10 Summary

Various measures of dispersion have been defined and formula for their calculation are given in this unit. Once data have been represented by a measure of central tendency, are may like to know the scatter of the given data around this measure of central tendency. The various measures of dispersion are range, quartile deviation, mean deviation, standard deviation and variance. Coefficient of variation for consistency of data or frequency distribution is defined as the ratio of standard deviation to arithmetic mean. It has no unit.

2.11 Further Readings

1. Goon, A.M., Gupta M.K. and B. Dasgupta : Fundamentals of Statistics, Volume I, the World Press Pvt. Limited, Calcutta.

2. Yule G.U. and Kendall M.G. : An Introduction to the theory of Statistics. Charles Griffin and Co. Limited.

NOTES

NOTES

U.P. Rajarshi Tandon Open University, Allahabad

UGSTAT-01
STATISTICAL METHODS

Block -III

Moments, Skewness and Kurtosis

Introduction

The **Block III** deals with moments, skewness and Kurtosis. It consist of two units.

Unit – 1 defines various moments of the frequency distribution and give the interrelationship between them.

Unit – 2 explains the significance of asymmetrical data and describes various measures of skewness i.e. lack of symmetry of data. It also gives the measures of Kurtosis and explains the peakedness of the frequency curve near the highest frequency.

Unit-1 : Moments, Raw Moments and Central Moments

Structure

1.1 Introduction

Measures of central tendency and variability (dispersion) enable us to know some important characteristic of the data and help us to compare two or more series. It can be illustrated that two different distribution may have the same mean and / or variance, still they may have different pattern of the distribution. Two other characteristics of the distribution are known as symmetry and peakedness of the curve. These may be defined in terms of the Central moments of the distribution.

The word moment is derived from statics, where moment about a point is equal to force multiplied by the perpendicular distance. The mean and variance of the distribution are the first moment about origin and second moment about mean or second central moment of the distribution.

1.2 Objectives

After going through this unit, you shall be able to –

- Compute raw moments including mean of the given data/frequency distribution.
- Compute central moments including variance of the given data/frequency distribution.
- Compute factorial moments of the given data/ frequency distribution.
- Use the interrelationship between these moments to obtain one from the other known moments.
- Apply charliers check and shephard's correction for moments.

1.3 Moments (Definition)

Suppose we have n values of a variable X as X_1, X_2, ...X_n. The possible measures of central tendency and dispersion of variable x are mean and variance defined by expression :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\overline{x}$ is the first moment of x about the origin.

1.3.1 Raw Moments for Ungrouped Data

Definition : If x_1, x_2, -----x_n are n values of the variable x, the r^{th} raw moment of x about any point A is defined as –

$$m_r' = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^r \; ; r = 0,1,2..... \qquad \ldots.(1.1)$$

So,

(always)

$$\left.\begin{aligned} m_0' &= \frac{1}{n}\sum (x_i - A)^0 = 1 \\ m_1' &= \frac{1}{n}\sum (x_i - A)^1 = (\bar{x} - A) \\ m_2' &= \frac{1}{n}\sum (x_i - A)^2 \\ m_3' &= \frac{1}{n}\sum (x_i - A)^3 \\ m_4' &= \frac{1}{n}\sum (x_i - A)^4 \end{aligned}\right\} \quad \text{.....(1.2)}$$

In particular, if the r^{th} raw data about origin, i.e. for A = 0 is

$$m_r' = \frac{1}{2}\sum x_i^r$$

so that,

$$m_0' = \frac{1}{n}\sum x_i^0 = 1 \,(Always)$$

$$m_1' = \frac{1}{n}\sum x_i = mean\ of\ the\ distribution$$

$$m_2' = \frac{1}{n}\sum x_i^2$$

$$m_3' = \frac{1}{n}\sum x_i^3$$

$$m_4' = \frac{1}{n}\sum x_i^4$$

1.3.2 Raw Moments for the grouped data

If the given values are in the form of a frequency distribution,

Table 1.1

Value of x_i x	x_1 x_2 .....x_i .....x_n
Frequency	f_1 f_2 ----------f_i .....f_n

the formula for moments about the point A takes the form

$$m_r' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^r; \quad r = 0, 1, 2 \quad \text{.....(1.3)}$$

Table 1.2

Class Interval	Mid-point of the class	frequency
	x_1	f_1
	x_2	f_2
	.	.
	.	.
	.	.
	.	.
	x_i	f_i
	.	.
	.	.
	.	.
	x_n	f_n
Total	--	$N=\sum fi$

We shall write it as "x_i/f_i (I = 1,2,....n) distribution"

where x_i is class mark of the i^{th} class, or its value of the variable X(Table 1.1), fi is its frequency and $N=\sum fi$ is total frequency. (number of observations)

If A = 0, m'_r is r^{th} raw moments about natural origin.

$$m'_0 = \frac{1}{N}\sum f_i x_i^0 = \frac{1}{N}\sum f_i = 1 \ (Always)$$

$$m'_1 = \frac{1}{N}\sum f_i x_i = mean\ of\ the\ distribution = \bar{x}$$

$$m'_2 = \frac{1}{N}\sum f_i x_i^2$$

$$m'_3 = \frac{1}{N}\sum f_i x_i^3$$

$$m'_4 = \frac{1}{N}\sum f_i x_i^4 \qquad (1.4)$$

1.3.3 Central Moments

If the arbitrary origin of moments of variable X is taken as arithmetic mean i.e. $A = \bar{x}$ the moments are called central moments.

Definition: For ungrouped data $x_1, x_2 \ldots\ldots\ldots x_n$ the r^{th} central moment of variable X is given by

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r;\ r = 0,1,2, \qquad \ldots\ldots\ldots(1.5)$$

If the given values are classified into a frequency distribution, x_i/f_i (i-1,2,…, n), the r^{th} central moment is given by –

$$m_r = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^r;\ r = 0,1,2, \qquad \ldots\ldots\ldots(1.6)$$

x_i being the mid-value of I^{th} or x^{th} value of the variable (as the case may be) class and f_i its frequency

Evidently, we have

$$m_0 = 1$$

and $m_1 = 0$ (Always) $\qquad \ldots\ldots\ldots\ldots(1.7)$

The second central moment of var iable X is variance of the distribution i.e

$$V(x) = m_2 = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2; & (\textit{for ungrouped data}) \\ \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^r; & (\textit{for grouped data}) \end{cases} \qquad \ldots\ldots(1.8)$$

Third and fourth central moments for ungrouped and grouped data are-

$$\begin{cases} m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 & (\textit{for ungrouped data}) \\ m_3 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^3 & (\textit{for grouped data}) \end{cases} \qquad \ldots\ldots(1.9)$$

$$\begin{cases} m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 & (\textit{for ungrouped data}) \\ m_4 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^4 & (\textit{for grouped data}) \end{cases} \qquad \ldots\ldots(1.10)$$

1.3.4 Factorial Moments

The r^{th} factorial moment about the origin of the distribution x_i/f_i (i=1, 2, 3,…… n) is defined as follows:

$$\mu_{(r)} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^{(r)}$$

where, $x^{(r)} = x(x-1)(x-2)....(x-r+1)$ *and* $N = \sum_{i=1}^{n} f_i$ .....(1.11)

Similarly, the factorial moment of order r about any point x=A is given as

$$\mu_{(r)} = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^{(r)}$$

where, $(x-A)^{(r)} = (x-A)(x-A-1)(x-A-2)...(x-A-r+1)$.....(1.12)

1.4 Inter-relationship between various moments

1.4.1 Central moments expressed in terms of raw moments

We have first r raw moments m'_1 , m'_2.... m'_r about an arbitrary origin A. The r^{th} order central moment can be obtained by using simple algebra.

We have

$$(x_i - \bar{x})^r = \{(x_i - A) - (\bar{x} - A)\}^r$$

$$= (x_i - A)^r - r_{C_1}(x_i - A)\}^{r-1}(\bar{x} - A) + r_{C_2}(x_i - A)\}^{r-2}(\bar{x} - A)^2$$

$$+ (-1)^r r_{C_r}(\bar{x} - A)^r \qquad .....(1.13)$$

Since

$$m'_1 = \frac{1}{N}\sum fi(x_i - A) = \bar{x} - A$$

The r^{th} central moment can be written as

$$m'_r = \frac{1}{N}\sum fi(x_i - \bar{x})^r$$

$$= \frac{1}{N}\sum fi(x_i - A)^r - r_{C_1}\frac{1}{N}\sum fi(x_i - A)^{r-1} m'_1 + r_{C_2}\frac{1}{N}\sum fi(x_i - A)^{r-2} m'^2_1$$

$$+(-1)^r m'^r_1$$

$$\therefore\ m_r = m'_r - r_{C_1} m'_{r-1} m'_1 + r_{C_2} m'_{r-2} m'^2_1 .....+(-1)^r m'^r_1 \qquad .....(1.14)$$

It is easily seen that (1.14) holds for moments obtained from ungrouped data as well.

Putting r = 1, 2, 3, and 4 in (1.14), we get some particular cases for A=0, as

$$m_1 = m'_1 - m'_1 = 0$$

$$m_2 = m'_2 - 2m'_1 m'_1 + m'^2_1$$

$$= m'_2 - m'^2_1$$

$$m_3 = m'_3 - 3m'_2 m_1 + 3m'_1 m_1^2 - m_1^3$$
$$= m'_3 - 3m'_2 m'_1 + 2m_1^3$$
$$m_4 = m'_4 - 4m'_3 m'_1 + 6m'_2 m_1^2 - 4m'_1 m_1^3 + m_1^4$$
$$= m'_4 + 4m'_3 m'_1 + 6m'_2 m_1^2 - 3m_1^4$$

.....(1.15)

In most of the practical problems, it is sufficient to calculate $\bar{x}$, m_2, m_3 and m_4 using calculators. These computations are greatly facilitated by first compiling moments about a suitably chosen origin A or origin 'O'. We first calculate

$$N = \sum f_i$$
$$\sum f_i x_i = \quad \sum f_i u_i =$$
$$\sum f_i x_i^2 = \quad \sum f_i u_i^2 =$$
$$\sum f_i x_i^3 = or \quad \sum f_i u_i^3 =$$
$$\sum f_i x_i^4 = \quad \sum f_i u_i^4 =$$

where, and then use (1.15) to compute $\bar{x}, m_2, m_3 m_4$ using relations (1.14) and (1.15).

For grouped data

$$u_i = \frac{x_i - A}{h} \qquad .....(1.16)$$

h is the common class interval, and A is a suitably chosen origin or reference point Theoretically, A can be any point , but it is so chosen that the computation work may be reduced. For grouped data, A is taken at a mid point near the central classes.

1.4.2 Raw Moments expressed in terms of central moments

Just as central moments can be expressed in terms of moments about an arbitrary origin A, so a moment about an arbitrary origin is expressible in terms of central moments.

From (1.2)

$m'_1 = \bar{x} - A$, and

$$m'_r = \frac{1}{N}\sum_i fi(x_i - A)^r$$

$$= \frac{1}{N}\sum_i fi(x_i - \bar{x} + \bar{x} - A)^r$$

$$= \frac{1}{N}\sum_i fi\{(x_i - \bar{x}) + m'_1\}^r$$

$$= \frac{1}{N}\sum_i fi\{(x_i - \bar{x})^r + r_{c_1}(x_i - \bar{x})^{r-1}m'_1 + r_{c_2}(x_i - \bar{x})^{r-2}m_1'^2 + m_1'^r$$

$$= \frac{1}{N}\sum_i fi\{(x_i - \bar{x})^r + r_{c_1}\frac{1}{N}\sum_i fi(x_i - \bar{x})^{r-1}m'_1 + r_{c_2}\sum_i fi(x_i - \bar{x})^{r-2}m_1'^2 + m_1'^r$$

$$= m_r + r_{c_1}m_{r-1}m'_1 + r_{c_2}m_{r-2}m_1'^2 + \ldots\ldots\ldots + m_1'^r \qquad \ldots\ldots\ldots\ldots\ldots(1.16)$$

In particular,

$$\mu_2 = \mu_2 + \mu_1'^2$$

$$\mu_3 = \mu_3 + 3\mu_2\mu_1 + \mu_1'^3$$

$$\mu_4 = \mu_4 + 4\mu_3\mu_1 + 6\mu_2\mu'^2 + \mu_1'^4$$

These formulae help us to obtain the moments about any point A, if the central moments are known.

1.5 Effect of change of origin and scale on central moments

If we change the origin of x on some arbitrary point A and scale by h. The new variable u is defined as $u = \frac{x - A}{h}$ so that $x = A + hu, \bar{x} = A + h\bar{u}$ and

$$m_r = \frac{1}{N}\sum_i fi(x_i - \bar{x})^r$$

$$= \frac{1}{N}\sum_i fi(hu_i - h\bar{u})^r$$

$$= h^r\frac{1}{N}\sum_i fi(u_i - \bar{u})^r$$

$$= h^r m_r(u) \qquad \ldots\ldots(1.17)$$

where $m_r^{(u)} = \frac{1}{N}\sum f_i(\mu_i - \bar{u})^r$

Thus, r^{th} central moment of variable X is h^r times r^{th} central moment of variable U. So, we conclude that central moment is unaffected by change of origin but it is affected by change of scale.

1.6 Charlier's Checks

Charlier's checks are often used as a ready check against some possible mistake in the calculation of first four moments. For this we first compute $\sum f_i x_i, \sum f_i x_i^2, \sum f_i x_i^3, \sum f_i x_i^4$ etc. and verify the calculations by the following identity.

$$\sum_i fi(x_i + 1)^4 = \sum_i fix_i^4 + 4\sum_i fix_i^3 + 6\sum_i fix_i^2 + 4\sum_i fix_i + N \quad \text{.....(1.18)}$$

1.7 Sheppard's Correction for moments

In computing moments for data grouped into class-intervals by means of the formulae

$$m'_r = \frac{1}{n}\sum_i x_i - A)^r f_i \text{ and } m_r = \frac{1}{n}\sum_i (x_i - \bar{x})^r f_i \quad \text{.....(1.19)}$$

In the computation of mean, variance and various moments from a grouped data into class intervals, we have taken mid-point or class mark as the representative of that class. Here, the assumption is that the mid-point of the classes are a reasonable approximations to the mean of the observations of that class. This approximation holds for good for the distribution which are symmetrical, moderately skewed and have classes having small classes intervals. It does not hold for all distributions.

We have the assumption that the observation falling in a class (e.g. the fi values falling in the ith class) were all equal to the class-mark or mid-point, although the observation may be really unequal. The assumption naturally introduces some error, which are called the errors due to grouping. To correct for these grouping errors, the computed values of the moments have to be suitably adjusted. A method for adjusting the moments for grouped data where the classes are equally wide has been developed by Sheppard. Sheppard's corrections for moments about an arbitrary origin and for central moments of the first four orders are given below:

$$m'_1(corrected) = m'_1$$

$$m'_2(corrected) = m'_2 - \frac{c^2}{12}$$

$$m'_3(corrected) = m'_3 - \frac{c^2}{4}m'_1,$$

$$m'_4(corrected) = m'_4 - \frac{c^2}{2}m'_2 + \frac{7}{240}c^4$$

and

$$m_2(corrected) = m_2 - \frac{c^2}{12}$$

$$m_3(corrected) = m_3$$

$$m_4(corrected) = m_4 - \frac{c^2}{2}m_2 + \frac{2}{240}c^4$$

where c is the width of each class-interval and m_r' s are the uncorrected rth moments for r = 1, 2, 3....

The sheppord's corrections is applicable only when –

(i) class-width are equal

(ii) the distributions are symmetrical or moderately skewed.

(iii) N is sufficiently large,

(iv) The frequency curve is not J or U shaped

(v) It is necessary that the observations should relate to a continuous variable.

1.8 Some Solved Examples

Example 1.1 : Find first four raw moments about x = 5 if the values of variables are 2, 3, 6, 8 and 11.

Solution:

$$m'_0 = 1$$

$$m'_1 = \frac{1}{n}\sum (x_i - A) = (\bar{x} - A)$$

$$\bar{x} = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6$$

$$m'_1 = 6 - 5 = 1$$

$$m'_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - A)^2$$

$$= \frac{(2-5)^2 + (3-5)^2 + (6-5)^2 + (8-5)^2 + (11-5)^2}{5}$$

$$= \frac{9+4+1+9+36}{5} = \frac{59}{5} = 11.8$$

Similarly,

$$m'_3 = \frac{(2-5)^3 + (3-5)^3 + (6-5)^3 + (8-5)^3 + (11-5)^3}{5}$$

$$= \frac{-27-8+1+27+216}{5} = \frac{209}{5} = 41.8$$

and

$$m'_4 = \frac{(2-5)^4 + (3-5)^4 + (6-5)^4 + (8-5)^4 + (11-5)^4}{5}$$

$$= 295$$

Hence, the first four raw moments are 1, 11.8, 41.8 and 295.

Example 1.2 : The number of suits sold daily by a women's boutique on the past six days has been given in the following frequency table.

Following frequency table.

Value (x)	Frequency (f)
3	2
4	1
5	3

Obtain first four raw moments about origin.

Solution :

r^{th} raw moments about origin is given by

$$m'_r = \frac{1}{N}\sum f_i x_i^r \qquad ; \qquad r = 0,1,2.....$$

$$m'_0 = 1$$

$$m'_1 = \frac{1}{N}\sum f_i x_i^r = \frac{(2\times 3)+(1\times 4)+(3\times 5)}{6}$$

$$= \frac{25}{6} = 4.17$$

Similarly,

$$m'_2 = \frac{2\times 3^2 + 1\times 4^2 + 3\times 5^2}{6}$$

$$= \frac{109}{6} = 18.17$$

$$m'_3 = \frac{2\times 3^3 + 1\times 4^3 + 3\times 5^3}{6}$$

$$= 82.17$$

$$m'_4 = \frac{2\times 3^4 + 1\times 4^4 + 3\times 5^4}{6}$$

$$= 382.17$$

Answer : First four raw moments are 4.17, 18.17, 82.17 and 382.17 suits respectively.

Example 1.3 : The following frequency table gives the values obtained in 15 throws of a die.

Value (x)	Frequency (f)
1	3
2	2
3	3
4	4
5	2
6	1

Find first four central moments.

Solution:

x_i	f_i	$f_i x_i$	$(x_i - \bar{x})$	$f_i(x_i - \bar{x})^2$	$\sum f_i(x_i - \bar{x})^3$	$\sum f_i(x_i - \bar{x})^4$
1	3	3	-2.2	14.52	-31.94	70.27
2	2	4	-0.2	2.88	-3.46	4.15
3	3	9	-0.2	0.12	-0.02	0.00
4	4	16	0.8	2.56	2.05	1.64
5	2	10	1.8	6.48	11.66	20.99
6	1	6	2.8	7.84	21.95	61.46
Total	N=15	48		34.40	0.24	158.51

$$\therefore \quad \bar{x} = \sum_{i=1}^{n} \frac{f_i x_i}{N} = \frac{48}{15} = 3.2$$

$$m_1 = 0$$

$$m_2 = \sum_{i=1}^{n} \frac{f_i (x_i - \bar{x})^2}{N} = \frac{34.40}{15} = 2.293$$

$$m_3 = \sum_{i=1}^{n} \frac{f_i (x_i - \bar{x})^3}{N} = \frac{0.24}{15} = 0.016$$

$$m_4 = \sum_{i=1}^{n} \frac{f_i (x_i - \bar{x})^4}{N} = \frac{158.51}{15} = 10.567$$

Answer: First four central moments are

$m_1 = 0$, $m_2 = 2.293$, $m_3 = 0.016$, $m_4 = 10.567$.

Example 1.4 : The first four raw moments of a distribution about the value 5 of variable are 2, 20, 40 and 50. Obtain mean and first four central moments.

Solution:

In the above example

$A = 5$, $m'_1 = 2$, $m'_2 = 20$, $m'_3 = 40$ and $m'_4 = 50$

$m'_1 = \bar{x} - A \Rightarrow m'_1 + A = 2 + 5 = 7 = \bar{x}$

Using relations in equations (2.9) we have

$$m_2 = m'_2 - m_1'^2 = 20 - 2^2 = 16$$

$$m_3 = m'_3 - 3m_2'^2 + 2m_1'^3$$

$$= 40 - 3 \times 20 \times 2 + 2 \times 2^3$$

$$= -64$$

$$m_4 = m'_4 - 4m'_3 m'_1 + 6m'_2 m_1'^2 - 3m_1'^4$$

$$= 50 - 4 \times 40 \times 2 + 6 \times 20 \times 2^2 - 3 \times 2^4$$

$$= 162$$

Answer: $\bar{x} = 7, m_2 = 16, m_3 = -64$ *and* $m_4 = 162$

Example 1.5 : Calculate the first four central moments of the following distribution.

x :	2	3	4	5	6	7	8	9
f :	1	8	28	56	70	28	8	1

We should shift the origin at a point say A to reduce the calculation. Let us choose A=5, (near the middle of the table). We may have taken

$A = \frac{5+6}{2} = 5.5$ but it will make the calculation more cumbersome.

Solution:

This is a symmetrical distribution with A=5 (near the middle of the table), and defining u=x-A, we get

$$\sum f_i u_i = 0 \text{ and } \sum f_i u_i^3 = 0$$

It has simplified the calculation –

Calculations

x	f	u = x-5	fu	fu^2	fu^3	fu^4
1	1	-4	-4	16	-64	256
2	8	-3	-24	72	-216	648
3	28	-2	-56	112	-224	448
4	56	-1	-56	56	-56	56
5	70	0	0	0	0	0

6	56	1	56	56	56	56
7	28	2	56	112	224	448
8	8	3	24	72	216	648
9	1	4	4	16	64	256
Total	256		0	512	0	2816

$$m'_1 = \sum \frac{fui}{N} = 0 \; ; \; m'_2 = \frac{\sum f_i u_i^2}{N} = \frac{512}{256} = 2$$

$$m'3 = \frac{\Sigma f_1 u_1^3}{N} = 0; \; m'_4 = \frac{\sum f_1 u_1^4}{N} = \frac{2816}{256} = 11$$

$$m_1 = 0$$

$$m_2 = m'_2 - m_1'^2 = 2$$

$$m_3 = m'_3 - 3m'_2 m'_1 + 2m_1'^3 = 0$$

$$m_4 = m'_4 - 4m'_3 m'_1 + 6m'_2 m'^2_1 - 3m'^4_1$$

$$= 11$$

Since central moments is unaffected by change of origin the first four central moments of variable X are for given distribution is

$$M_i(x) = A + M_1'^{(u)} = 5 \; m_2 = 2, \quad m_3 = 0 \quad m_4 = 11$$

You may try the following problems.

1.9 Summary

In this unit the properties of skewness and kurtosis of a distribution are discussed. The departure of the distribution from symmetry is known as skewness. The convexity of the curve is called kurtosis. It is compared with the peakedness of the normal distribution. Various measures of skewness and kurtosis are also discussed.

1.10 Exercises

P-1.1 Define raw and central moments of a frequency distribution. Obtain the relationship between the central moments of order r in terms of the raw moments. What is Sheppard Corrections to the central moments?

P-1.2 Define moments. Express central moments in terms of raw moment and vice-versa.

P-1.3 The first three moments of a distribution about the value 2 of the variable are 1, 16 and –40. Show that mean is 3, the variance is 15 and μ_3=-86.

P-1.4 Express first four central moments in terms of raw moments. What is the effect of change of origin and scale on central moments?

P-1.5 The central moments of a frequency distribution of incomes of a labour class are $m_2 = 100\ Rs^2$, $m_3 = -4\ Rs^3$ and $m_4 = 624\ Rs^4$.

(a) If the income of each labour is increased by 10 Rs, what will be the moments?

(b) If the income of each labour is doubled. What will be new moments?

P-1.6 The first three moments of a distribution about the value x = 7 are 3, 10 and 15 respectively. Obtain mean, variance and m_3.(Ans. 10, 1, -2.1).

P-1.7 The first four raw moments of a distribution about x = 4 are 1, 4, 10, 45. Show that the mean is 5 and the variance is 3 and m_3 and m_4 are 0 and 26 respectively.

P-1.8 What is Sheppard's Correction ? What will be the corrections for the first four raw moments and first four central moments?

P-1.9 Find the second, third and fourth central moments for the frequency distribution given below :

Class Interval	Frequency
110-115	05
115-120	15
120-125	20
125-130	35
130-135	10
135-140	10
140-145	05

Also apply Sheppard Correction for moments.

P-1.10 Frequency distribution of scores in mathematics of 50 students are given below:

Score	50-60	60-70	70-80	80-90	90-100
Frequency	1	0	0	1	1
Score	100-110	110-120	120-130	130-140	140-150
Frequency	2	1	0	4	4
Score	150-160	160-170	170-180	180-190-	190-200
Frequency	2	5	10	11	4
Score	200-210	210-220	220-230		
Frequency	1	1	2		

Compute first four central moments. Obtain corrected central moments after applying the Sheppard corrections.

(Answer: $m_2 = 1,176$; $m_3 = -41,160$; $m_4 = 57,45,600$;

Corrected moments are $m_2 = 1167.67$; $m_3 = -41160$ and $m_4 = 5687091$.

1.11 Answers and Suggestions

P- 1.5 (a) No change m, m_2, m_3, m_4

(b) Here h=2, therefore, $m_2 = 400\ Rs^2$, $m_3 = -32\ Rs^3$ $m_4 = 9984\ Rs^4$

P – 1.6 $\mu_1 = 10$, $\mu_2 = 10-3^2 = 1$, $\mu_3 = 15-3 \times 10 \times 3 + 2 \times 3^3 = -21$

P – 1.9 Solution is shown below:

Class Interval	Mid-point x	$\mu_i = \frac{x_i 127.5}{5}$	Frequency f_i	$f_i\mu_i$	$f_i\mu_i^2$	$f_i\mu_i^3$	$f_i\mu_i^4$
110-115	112.5	-3	05	-15	+45	-135	405
115-120	117.5	-2	15	-30	+60	-120	240
120-125	122.5	-1	20	-20	+20	-20	+20
125-130	127.5	0	35	0	0	0	0
130-135	132.5	1	10	10	10	1-	10
135-140	137.5	2	10	20	40	80	160
140-145	142.5	3	05	15	45	135	405
Total			N=100	-20	220	-50	1240

For distribution for μ

Raw moments

$$\mu_1'(\mu) = \frac{-20}{100} = -0.2, \qquad \mu_2'(\mu) = \frac{220}{100} = +2.20$$

$$\mu_1'(\mu) = \frac{-50}{100} = -0.5, \qquad \mu_2'(\mu) = \frac{1240}{100} = 12.40$$

Central moments

$$\mu_2(\mu) = \mu_1'^2(\mu) = 2.20 - (-0.2)^2 = 2.20 - 0.04 = 2.16$$

$$\mu_3(\mu) = \mu_3(\mu) - 3\mu_2(\mu)\mu_1'(\mu) + 2\mu_1'(\mu) = -0.5 - 3 \times 2.20 \times (-0.2) + 2(-0.2)^3$$

$$= -0.5 + 1.32 - 0.016 = 0.804$$

Similarly,

$$\mu_4(\mu) = 12.40 - 4 \times (-0.5)(-0.2) + 6(2.20)(-0.2)^2 - 3(-0.2)^4$$

$$= 12.40 - 0.4 + 0.528 - 0.0048 = 12.5232$$

Hence, for the given distribution , with h=5.

$$\mu_2 = 5^2 \times 2.16 = 54.0$$

$$\mu_3 = 5^3 \times 0.804 = 100.5$$

$$\mu_4 = 5^4 \times 12.5232 = 7827.0$$

Sheppard correction.

$$\bar{\mu}_2 = m_2 - \frac{h^2}{12} = 54 - \frac{5^2}{12} = 54.02 = 51.9167$$

$$\bar{\mu}_3 = m_3 \qquad = 100.50$$

$$\bar{\mu}_4 = m_4 - \frac{h^2}{2} m_2 + \frac{2}{240} h^2 = 7827 - \frac{25}{2} \times 54 \times \frac{2}{240} \times 5^4 = 7157.2083$$

1.12 Further Readings

1. Goon, A.M., Gupta M.K. and B. Dasgupta : Fundamentals of Statistics, Volume I, the World Press Pvt. Limited, Calcutta.

2. Yule G.U. and Kendall M.G. : An Introduction to the theory of Statistics, Charles Griffin and Co. Limited.

Unit-2 Skewness and Kurtosis

Structure

2.1 Introduction

To have an idea about the shape of the frequency or probability curve we study the skewness and kurtosis of the distribution. A distribution is said to be symmetrical if the frequencies (probabilities) are equal on either side of the central value. It implies that both the right and left tails of the curve are exactly equal in shape and length. If a distribution is not symmetrical then it is called asymmetric or skewed in the direction of the extreme values, i.e., on the right – or on the left. Since extreme values give longer tail in its direction therefore, the distribution having longer right-tail is called right skewed or positively skewed distribution. The left implies longer left tail. Thus a measure of skewness indicates the extent as well as direction of skewness of the distribution.

Karl Pearson called a normal curve as mesokurtic, it has a hump at the middle. Pearson defined Kurtosis as the convexity of the curve and used β_2 and γ_2 as its measure. The measure of Kurtosis gives an idea whether the center of the distribution is assuming flatness or peakedness similar to the hump of the normal probability curve or not.

The measures of skewness are very useful in biological, chemical and physical laboratory works. They are used in economic social statistics and medical statistics to study the behaviour of the data.

2.2 Objectives

After going through this unit, you shall be able to :

- Differentiate between the behaviour of symmetrical data and right or left skewed data.
- Obtain the measures of skewness and kurtosis and interpret them.

2.3 Skewness and its measures

Definition : By skewness of a frequency distribution we mean the degree of its departure from symmetry. The frequency distribution of a discrete variable x is called symmetrical about the value x_0 if the frequency of x_0-h is the same as the frequency of x_0+h, whatever h may be.

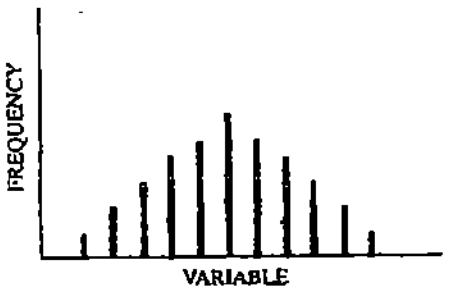


Fig. 2.1a A symmetrical distribution (discrete variable).

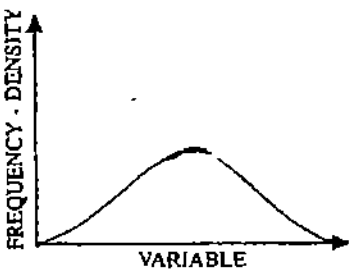


Fig. 2.1b A symmetrical distribution (continuous variable).

In the case of a continuous variable, the term 'symmetry' should be used in relation to its frequency curve. The frequency curve of a continuous variable is said to be symmetrical about x_0 if the frequency density at x_0–h, is

the same as the frequency-density at $x_0 + h$, whatever h may be. Figures 2.1a and 2.1b show two symmetrical distribution.

A distribution which is not symmetrical is called asymmetrical or skew. This skewness is said to be positive if the longer tail of the distribution is towards the higher values of the variable (Fig.2.2a), and negative if the longer tails is towards the lower vales of the variable (Fig. 2.2b)

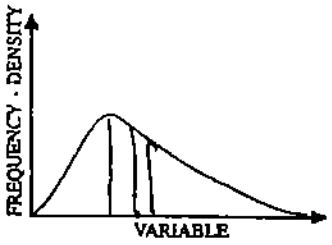


Fig. 2.2a A positively skew distribution.

Mode < Median < Mean

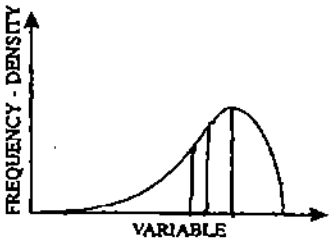


Fig. 2.2b A negatively skew distribution.

Mode < Median < Mean

An important point to be noted in this connection is that all odd-order central moments are zero for a symmetrical distribution, positive for a positively skew distribution and negative for a negatively skew distribution. Any such moment may, therefore, be considered a measure of the skewness of a distribution except, of course, m_1 which is necessarily zero for any distribution-symmetrical or otherwise. The simplest of these measures is m_3.

2.4 Measures of Skewness

2.4.1 Pearsons Coefficient

An alternative measure of skewness is obtained from the relative positions of the mean and the mode in a distribution.

In a symmetrical distribution, the mean, median and mode (assuming the distribution to be uni-modal) coincide. If the distribution is : skewed positively, then

Mean > median > mode.

and if it is negatively skewed, then

mean < median < mode.

Hence the difference (mean mode) divided by the s.d., is taken as a measure of skewness.

$$Sk = \frac{\bar{x} - Mo}{s} \qquad \text{.....(2.3)}$$

This is known as Pearson's first measure of skewness, provided s > 0.

Since it is difficult to estimate the mode from a frequency distribution, the empirical relation is used to get another measure of skewness viz.

The moments are not unit free quantity to make it unit free Karl Pearson defined the following four coefficients based on first four central moments.

$$\beta_1 = \frac{m_3^2}{m_2^3}, \beta_2 = \frac{m_4}{m_2^2}, \gamma_1 = \sqrt{\beta_1} \text{ and } \gamma_2 = \beta_2 - 3 \qquad \text{.....(2.1)}$$

To get the unit free measurement of skewness is thus defined as

$$g_1 = \frac{m_3}{m_2^{3/2}} \qquad \text{.....(2.2)}$$

$\gamma_1 = |g_1|$ is absolute measure of skewness.

An alternative measure of skewness is obtained from the relative positions of the mean and the mode in a distribution.

In a symmetrical distribution, the mean, median and mode (assuming the distribution to be uni-modal) coincide. If the distribution is positively skew, then

mean > median > mode.

and if it is negatively skew, then

mean < median < mode.

Hence the difference (mean-mode) divided by the s.d. is taken as a measure of skewness :

$$Sk = \frac{\bar{x} - Mo}{s} \qquad \text{.....(2.3)}$$

This is known as Pearson's first measure of skewness, provided s > 0

Since it is difficult to estimate the mode from a frequency distribution, the empirical relation is used to get another measure of skewness, viz.

$$Sk = \frac{3(\bar{x} - Mi)}{s} \qquad \text{.....(2.4)}$$

which is known as Pearson's second measure of skewness. If mean = Median = Mode then Sk = 0.

2.4.2 Bowley's Coefficient

The measure (2.4) can vary between --3 and 3. The same may be said to be approximately the case with (2.3) because of the empirical relation which is valid for moderately skew distribution. A fourth measure of skewness is obtained by considering the relative positions of the three quartiles of a frequency distribution. For a symmetrical distribution the lower and upper quartiles are equidistant from the median; for a positively skew distribution the lower quartile is nearer the median than the upper quartile is, while for a negatively skew distribution the upper quartile is nearer.

Thus $(Q_3 - Mi) - (Mi - Q_1)$ may be taken as a measure of skewness. It s expressed as a pure number on being divided by

$$(Q_3 - Mi) + (Mi - Q_1) = Q_3 - Q_1$$

which is assumed to be non-zero. Thus the new measure is

$$Sk = \frac{(Q_3 - Mi) - (Mi - Q_1)}{Q_3 - Q_1} \qquad \text{.....(2.5)}$$

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

This is known as Bowley's measure of skewness. As regards (2.5) , it has the limit --1 and 1.

2.4.3 β and γ

The moments are not unit free quantity to make it unit free Karl Pearson defined the following four coefficients based on first four central moments.

$$\beta_1 = \frac{m_3^2}{m_2^3}, \beta_2 = \frac{m_4}{m_2^2}, \gamma_1 = \sqrt{\beta_1} \text{ and } \gamma_2 = \beta_2 - 3 \qquad \text{.....(2.1)}$$

To get the unit free measurement of skewness is thus defined as

$$g_1 = \frac{m_3}{m_2^{3/2}} \qquad \text{.....(2.2)}$$

$\gamma_1 = |g_1|$ is absolute measure of skewness. The measure given by equation (2.2) can theoretically assume any value between $-\infty$ and ∞ but in practice its numerical value is rarely very high. For symmetrical distribution $\beta_1 = 0$. $\beta_1 > 0$ $\Rightarrow$ distribution is +1 relysk skewed. $\beta_2 < 0 \Rightarrow$ distribution is –1 rely skewed.

2.4.4 Another measure based on moments

May be obtained from the Pearson's system of curves. It is defined as –

$$S_k = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} \qquad \text{.....(2.6)}$$

It is without sign, It has a drawback that it has no limits. If $S_K = 0$ then $\beta_1 = 0$ or $\beta_2 = -3$.

$\beta_2 \neq -3$. Since $\beta_2 = \mu_4/\mu_2^2$. Hence $S_K = 0$, if f. $\beta_1 = 0$ Thus, for a symmetrical μ_2^2 distribution, $\beta_1 = 0$.

2.5 Kurtosis

Another method of describing a frequency distribution is to specify its degree of peaked-ness or kurtosis. Two distribution may have the same mean and the same standard deviation and may be equally skew, but one of them may be more peaked than the other.

2.5.1 Measure of Kurtosis β_2 and γ_2

This feature of the frequency distribution is measured by

$$\beta_2 = \frac{m_4}{m_2^2} \text{ and}$$

$$\gamma_2 = \beta_2 - 3 \qquad \text{.....(2.7)}$$

Obviously, it is a pure number. For a normal distribution, $\beta_2=3$ and $\gamma_2=0$. A positive value of γ_2 indicates that the distribution has high concentration of value near the central tendency and has high tails, in comparison with a normal distribution with the same standard deviation. In the same way, a negative value γ_2 means that the distribution has low.

$\beta_2 = 3$ implies that $\gamma_2 = 0$. the Kurtosis is same as that of normal curve. The curve is mesokurtic.

$\beta_2 > 3 \Rightarrow \gamma_2 > 0$, the Kurtosis is said to be positive and the curve callec the lepto Kurtosis.

$\beta_2 < 3 \Rightarrow \gamma_2 < 0$, the Kurtosis is said to negative and the curve is called platy kurtic.

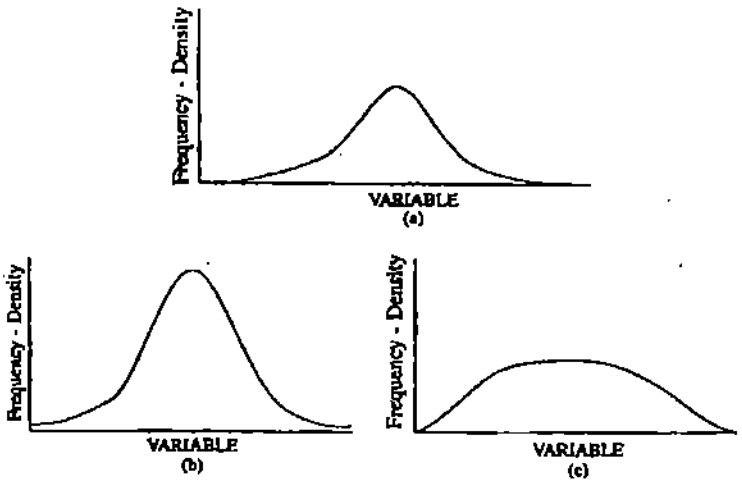


Fig. Three symmetrical distributions with different degrees of kurtosis : (a) mesokurtic, (b) leptokurtic, (c) platykurtic.

concentration of values in the neighbourhood of the central tendency and low tails, compared to a normal distribution with the same standard deviation. A normal curve is said to be mesokurtic (i.e. having medium kurtosis). A distribution with positive γ_2 is called leptokurtic, and one with negative γ_2 is known as platykurtic. The quantities β_1 and β_2 themselves are sometimes used as measures of skewness and kurtosis, respectively.

That the fourth central moment (m_4) may be used in measuring kurtosis becomes obvious from the fact that the higher the kurtosis, the higher will be the effect of the large deviations (from the mean) in the tails when raised to the fourth power. Division of m_4 by s^4 makes the measure a pure number.

Actually, however, β_2 (or γ_2) will be appropriate as a measure of kurtosis or peaked-ness only if we confine our attention to the class of the usual bell-shaped (or unimodal) distributions. Otherwise, it may only serve to distinguish a unimodal distribution from a bimodal.

Some Solved Examples

Example 2.1 In a frequency distribution $Q_1 = 30$, $Q_3 = 70$ and median is 38.

Compute coefficient of skewness.

Solution :

Coefficient of skewness based on quantities is

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

$$= \frac{70 + 30 - 2 \times 38}{70 - 30}$$

$$= 0.6$$

Ans : Coefficient of skewness is 0.6

Example 2.2 Mean, median, mode and standard deviation of frequency distribution are 25, 27, 35 and 15 respectively. Compute coefficients of skewness based on mean, mode, median and standard deviation.

Solution :

Coefficient of skewness based on mean, mode and S.D. is

$$Sk = \frac{\bar{x} - Mo}{s} = \frac{25 - 35}{15} = \frac{-10}{15} = -0.67$$

coefficient of skewness based on mean, median and S.D. is

$$S'k = \frac{3(\bar{x} - Mi)}{s} = \frac{3(25 - 27)}{15} = \frac{-6}{15} = -0.4$$

Ans : coefficient are skewness are -0.67 an -0.4

Example 2.3 First there central moments of a variable are 0,16 and -64 respectively. Compute coefficient of skewness based on moments and comment.

Solution :

Coefficient of skewness based on moments are :

$$\beta_1 = \frac{m_3^2}{m_2^3} = \frac{(-64)^2}{16^3} = 1$$

$$\gamma_1 = \sqrt{\beta_1} = 1$$

Since m_3 is negative, the distribution is negatively skewed, coefficient of skewness with sign is

$$g_1 = \frac{m_3}{s^3} = -1$$

Example 2.4 The standard deviation of a symmetrical distribution is 5 and fourth central moment is 2000. Compute β_2 and γ_2 and comment on the kurtosis of the distribution

Solution :

Since distribution is symmetric

T $m_3 = 0,\ m_2 = (\text{S.D})^2 = 25,\ m_4 = 2000$

then $$\beta_2 = \frac{2000}{25^2} = \frac{2000}{625} = 3.2 > 3$$

$$\gamma_2 = \beta_2 - 3 = 0.2 > 0$$

Since $\beta_2 > 3$ and $\gamma_2 > 0$; the distribution is leptokurtic.

2.6 Exercises

2.1 What are skewness and kurtosis ? Give some suitable measures for skewness and kurtosis.

2.2 Using Cauchy-Schwarz inequality, or otherwise, prove that

(i) $\beta_2 \geq 1$ and (ii) $\beta_2 - \beta_1 - 1 \geq 0$.

2.3 Show that the measure of skewness given by (2.4) must lie between-3 and 3 and that the measuring given by (2.5) must be between-1 and 1.

2.4 Prove, by a geometrical argument. that for a J-shaped distribution with its longer tail towards the higher values of the variable, the median is nearer to the first quartile than to the third (A similar argument can be used to show that for the other type of J-shaped distribution, the median is nearer to the third quartile than to the first).

2.5 Consider any symmetrical frequency distribution for a discrete variable. Show that its central moments of odd orders must all be zero.

2.6 The first three moments about 4 of 10 observations were 5.5, 38.5 and 302.5. The 4^{th} moment about 2 of the same 10 observations was 6089.3 It was found later that an observation of 3 was wrongly read as 8. Find the corrected mean, second, third and fourth central moments and measures of skewness and kurtosis.

2.7 In a certain distribution,

mean = 45 units,

median = 48 units.

coefficient of skewness = 0.4.

The person who supplied the data failed to give the value of the s.d. Estimate it from the above data.

2.8 In a certain distribution the coefficient of skewness based on quartiles is 0.6 if the sum of the third and first quartile is 100 units and the median is 38 units, find the first and third quartiles.

2.9 Compute $\bar{x}, s, m_3$ *and* m_4 for the data on length of ear-hear given in Exercise 6.16.

2.10 The scores in English of 250 candidates appearing at an examination have

mean = 93.72, m_2 = 97.80, m_3 = -114.18 and m_4 = 28.396.14.

It is later found on scrutiny that the score 61 of a candidate has been wrongly recorded as 51. Make necessary corrections in the given values of the mean and the central moments.

2.11 Particulars relating to the monthly wage distribution of two manufacturing firms are given below :

	Firm A	Firms
Mean wage	Rs. 1,477	Rs. 1,495
Median wage	Rs. 1,389	Rs. 1.354
Modal wage	Rs. 1,350	Rs. 1..312
Quartiles	Rs. 1,278 and 1,422	Rs. 1,262 and 1,435
Standard deviation	Rs. 87	Rs. 99

Compare the two distributions.

2.12 The S.D. of a symmetrical distribution is 4. What must be the value of fourth moment about mean so that the distribution be (a) leptokurtic (b) mesokurtic and (c) platy kutic.

2.7 Answers / Suggestions

P-2.6 Partial answers : Corrected first and second moments are 0 and 346.

P-2.7 22.5 unit

P-2.8 30 units

P-2.9 $\bar{x} = 9.9, s = 0.91, m_3 = -0.061, and\ m_4 = 29165.60$

(in proper units)

P-2.10 mean = 39.76, $m_2 = 99.10$, $m_3 = -93.27$ and $m_4 = 29165.60$

P-2.11

P-2.12 (a) $\mu_4 < 756$ (b) $\mu_4 = 756$ (c) $\mu_4 > 756$

2.8 Summary

In this unit the properties of symmetrical and asymmetrical distributions have been studies. The measures of skewness have been defined and their interpretations have been given. The peakedness of the frequency curve is explained . The measures of peakedness have been obtained in terms of Central moments.

2.9 Further Readings

1. Kenney, J.F and Keeping, E.S. : Mathematics of Statistics, Part I (Ch.7) Van Nostrand, 1954, and Affiliated East-West Press.
2. Mills. F.C. : Statistical Methods (Ch.5) H. Holt, 1955.
3. Yule, G.U. and Kendall, M.G. : Introduction to the Theory of Statistics (Ch.6) Charles Griffin, 1953.
4. Fundamental of Statistics Vol.I by Goon, Gupta Dasgupta.

NOTES

NOTES

NOTES