PREFACE

In the curricular structure introduced by this University for students of Post-Graduate degree programme, the opportunity to pursue Post-Graduate course in Subject introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with the objectives of open education which does not believe in artificial differentiation.

Keeping this in view, study materials of the Post-Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analysis.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing and devising of proper lay-out of the meterials. Practically speaking, their role amounts to an involvement in invisible teaching. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by the other.

The more a learner would seriously pursue these study materials the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that they may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the University.

Needless to add, a great part of these efforts is still experimental–in fact, pioneering in certain areas. Naturally, there is every possibility of some lapse or deficiency here and there. However, these to admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

Professor (Dr.) Ranjan Chakrabarti
Vice-Chancellor

NETAJI SUBHAS OPEN UNIVERSITY

Post-Graduate Degree Programme

Master of Business Administration (MBA)

Course Code : CP-104

Course Title : Statistics For Management

First Print : March, 2023

Printed in accordance with the regulations of the
Distance Education Bureau of the University Grants Commission.

NETAJI SUBHAS OPEN UNIVERSITY

Post-Graduate Degree Programme
Master of Business Administration (MBA)
Course Code : CP-104
Course Title : Statistics For Management

: Course Writer :
Prof. Nishit Kr. Patra

: Course Editor :
Prof. Ajoy Kumar Biswas

: Format Editor :
Professor Anirban Ghosh
Director (i/c), (Chairperson)
School of Professional Studies, NSOU

Netaji Subhas Open University

Master of Business Administrate (MBA)

Course Code : CP-104

Course Title : Statistics For Management

Unit 1 ❑ Statistics and its representation

Structure

1.0 Objectives

Every person requires data on what in going on is every sphere of his or her life regarding livelihood, culture, education etc., in the society, where he or she lives. For this, statistics gives him or her information regaring the answers to his or her, queries. Ihe word Statistics conveys a variety of meanings. To some people it is part of Mathematics and to some others it is collection of tables, charts and figures, which are commonly seen in books, journals and newspapers. For this we have to define statistics properly and give its different functions like applications, limitations etc., and also classification and representation of data in this chapter.

1.1 Introduction

The word 'Statistics' came from the Latin word 'Status' and is used for the informations relating to state. Old concept of this was that it gave various informations about a state for some administrative purpose in the fields of Agriculture, Industry, Health, Games etc. Statistics is the collection of data i.e., informations collected from a group of individuals and is also a scientific method of representation, analysis and interpretation of the dataset. Statistics is regarded as one of the most important tools for taking decisions in the midst of uncertainties. Since Statistics helps to form suitable policies, need of it is required in almost all sphere of life. Informations collected either from the field of enquiry or from some records, where informations are already collected, are required in solving different problems in our day to day life and in drawing inferences there.

1.2 Statistical Methods

By 'data' we mean collection of observations obtained from a field of (statistical) investigation directly or from some publications. Data is always used

in the collective sense but not in singular sense. For example, total scores of students, appeared in Higher Secondary Examination 2004 in West Bengal, is the data (or data set).

Statistics, in the singular sense, is the body of methods that are used for the treatment of data like collection, organisation, presentation, analysis and interpretation, and is the collection of data arising due to some investigation or collection from some publication in the plural sense.

Five stages of statistical investigation of data are

(i) **Collection :** Utmost care must be taken to collect data because they are the foundation of statistical analysis, otherwise for faulty data the conclusions drawn will be false.

(ii) **Organisation :** Data collected from publications are in organised form. But when data are collected from an investigation they should be organised by editing them very carefully so that omissions, inconsistencies, irrelevant answers and wrong computation should be corrected or adjusted properly. After editing the data they are arranged according to some common characteristics possessed by the items constituting the data.

(iii) **Presentation :** Then the data should be tabulated in common rows or columns according to common characteristics of them so that data will be presented with clarity. They can also be presented in diagrams or graphs to facilitate statistical analysis.

(iv) **Analysis :** After collection, organisation and presentation of data they are analysed by using measures of different characteristics of data, namely, central tendency, dispersion, skewness, kurtosis, correlation, regression etc.

(v) **Interpretation :** From the data collected and analysed valid conclusions of study are drawn from correct interpretation of data to take valid decisions. Otherwise the whole objective of the investigation will be vitiated.

Functions of statistics are definitness, condensation, comparison of data, formulating a hypothesis and testing it by the help of a sample drawn from whole set of data, prediction and formulation of policies. Statistics presents the fact in a precise and definite form to comprehend the statements properly. The

facts conveyed in exact quantitative term are more convincing than description. Statistical method condeuse the mass of data to give meaningful overall information. For example, average per capita income is used to know the income position of a country. Statistics enables better appreciation of significances of conclusions drawn from a series of data by suitable device for comparison of them. Statistical methods are helpful in formulating a hypothesis and testing it to get suitable conclusions. Knowledge of future trends are very helpful in formulating suitable policies and plans. These future trends can be furecasted by suitable statistical methods. Statistics provides basic material for forming suitable policies.

1.2.1 Functions of Statistics

Statistics has universal applicability. All human activities are connected with statistical data. The functions of statistics are the following :

(a) **It simplifies complexity** : Human cannot understand a large number of facts and figures at a time. So, an important function of statistical methods is to simplify the complex data into diagrammatic representations, averages etc. which are easier to understand.

(b) **It is precise and definite** : It presents statements in a precise and definite form. Numerically expressed conclusions are more convincing than any other method, as they are precise and definite. For this, it is popular in various sciences.

(c) **It helps comparison** : It helps to compare one phenomenon with other or present results with past, because people are interested in relative figures than absolute figures in case of comparison.

(d) **It enlarges individual experiences** : One can easily grasp ideas from a condensed form of statistics, converted from mass by individual experiences. Statistics enables one to understand clear and definite ideas.

(e) **It formulates and tests hypothesis** : It is helpful in formulating and testing hypothesis. It provides guidance in the formulation of new policies and theories at all stages and in the drawing of plans in all fields.

(f) **It tests the laws of other sciences** : It helps in testing the laws of physical sciences and social sciences. Its techniques help hypothesis to

become a law by testing its truthfulness.

(g) **It for sees future courses :** Plans and policies can be formulated quite well in advance of time of implementation by using some statistical theories. A knowledge of future trends is very helpful in framing suitable policies and plans.

(h) **It helps to study relationship between different data :** Coefficient of correlation, coefficient of association, regression etc. are the measures through which functional relationship between two sets of data can be measured.

(i) **It helps the Government :** It is essential for the proper administration of a country. It provides informations needed for the efficient conduct of government business. Govt. uses statistics to have an understanding before implementing schemes like old age pensions, welfare schemes etc. The use of statistical data and techniques is inevitable in almost all ministries and departments of government. Thus it is of great use in the affairs of the state.

1.2.2 Scope of Statistics

Statistics is used as a key tool in the following fields.

(1) **States :** Statistics is heavily used in ruling personalities and in framing suitable policies, millitary or fiscal etc. The state collects statistics to solve several problems and these statistics help government to frame suitable policies in promoting human welfare.

(2) **Business and commerce :** In expanding competitive market, the problems relating to business and commerce become complex and so more statistical methods are used in decision making. Most of productions are in anticipation of demand and for this very careful study of market is required. Business runs on estimates and probabilities. Higher the degree of accuracy of a business man's estimates, the greater is the success attending on his business. To take decisions in policies of business and commerce, statistical methods are of great help.

(3) **Economics :** Economic deals with production and distribution of wealth with the complex institutional set up. Concerned with consumption, saving and investment of income, statistical methods are of great help in the proper undertaking of the economic problems and in the formulation of economic

policies. Econometrics considered as application of statistical methods is widely used in the field of research in economics.

(4) **Science and other fields :** Statistical methods, probabilistic methods, principle of uncertenty, sampling techfiiques are commonly used in different spheres of science and technology. In political science international relationship is also determined through the use of statistics.

1.3 Limitations of Statistics

Limitations of Statistics are the following :—

Statistics does not deal with individual but with aggregate, since statistics are aggregates of facts.

Statistics are numerical statements of facts since non-numerical expressions are incapable of statistical analysis.

Statistical results are true on an average. The conclusions drawn under statistical analysis are not universally true, but true under certain conditions.

Statistics is only one of the methods of studying a problem. Statistical tools do not provide the best solutions under all circumstances.

Statistics can be misused due to several reasons. For example, statistical conclusions based on incomplete information may give falacious conclusions. This requires experience and skill to draw suitable conclusions from the data, otherwise wrong interpretation may come.

Statistics cannot be used properly in the absence of proper understanding of the subject to which it is applied.

1.4 The collection of data

According to the basic source of data, data may be classified into two types : Primary data and Secondary data.

Primary data are those data which are collected first time directly from

the field of enquiry for a specific purpose. Here data are original and used by the collecting authority. Examples of primary data are data in Annual report of Railway Board published by the Ministry of Railway, Govt. of India and the data in Reserve Bank of India Bulletin, published by Reserve Bank of India.

Secondary data are those data which have already been collected by some agency or organisation for a specific purpose and are compiled from some suitable source for application in different connection. Here data are used by other agencies than the collecting authority of original data, who are responsible for the publication of the data. Examples of secondary data are the data in Statistical Abstract of India (Annual) published by Central Statistical Organisation, Govt. of India and the data in Annual statement of foreign trade, published by DGCIS, calcutta.

Relative advantages and disadvantages of primary data over the secondary data are the following.

In secondary data some information may be suppressed or condensed, but in primary data detailed information can be obtained.

Primary data is free from transcribing errors and estimation errors, whereas secondary data may contain such errors.

In primary data precise definition of the terms used and the scope of the data can be explicitly stated whereas there may be no mention of them in secondary data.

Primary data include the method of procuring data or approximations used whereas secondary data does not include them.

Vital point for using secondary data is cost effectiveness. Thus time, cost, suitability and accuracy are the essential factors to decide whether primary data or secondary data would be used.

Collection of data could be done either by complete enumeration or by sampling. In former case, each and every indidividual of the group, to which data are considered, is covered and informations are gathered for each individual separately. In the latter case only some individuals forming a representative part of the group are covered, either because the group is too large or because the items on which the information is sought are numerous. Complete enumeration

may lead to greater accuracy and more refinement in analysis, but it may be very expensive and time consuming. A sample, designed and drawn with care, can produce sufficiently accurate results for the purpose of enquiry and it can save much time, money and labour.

1.4.1 Methods of collection of Primary data

Primary data may be collected by investigators in following way :—

(i) **By direct personal observation :** An investigator(s) personally collects data from field of enquiry. He meets and interrogates persons, capable to supply the information from the field. He must be keen observer, tactful and polite in behariour to collect the data. The data thus obtained from the field of enquiry are quite reliable and accurate, but at the same time costly, time consuming. Because of personal approach the response of the data would be good and uniformity and homogeneity in data can be maintained. Misinterpretation, if any, on the part of information can be avoided. This method is required when greater accuracy, confidential data and intensive study of data is needed. Then field of enquiry should be small and sufficient time is available.

(ii) **By indirect oral investigation :** When the informant is reluctant to supply information, this method is followed by approaching the witness or third parties who are in touch with the informant. This method is simple and convenient. It save time, money and labour. It is used in investigation of large areas and in getting adequate information directly. Generally this method is employed by different enquiry commissions or committees and police departments by interrogating third parties having knowledge about the happenings of the fact under study. Drinking, gambling and other bad habits of people can also be known by this method. Interviews with improper men can spoil the result. Witness may colour the information according to their interests. So to get real position, sufficient number of persons are to be interviewed.

(iii) **By sending questionnaires by mail :** A questionnaire is a proforma containing relevant questions required for the object of enquiry. Questionnaires are sent to selected persons by mail with a request to return them duly filled up in return post. This method is economical and saves labour and time. This method is used to cover large area of investigation. This method is suitable only for the literate people. As there is no contact between investigator and

respondent, accuracy, reliability of the data may be hampered.

(iv) **By sending schedules through paid trained investigators :** This is very popular and effective method particularly in market research. Schedules are prepared and investigators are appointed and trained to meet the people concerned and to ask them accordingly to get required information for filling up of the schedules on spot on the basis of answers of the informants. The success of this method depends on the tactfulness, personality and intelligence of the investigators. By this method responsive, reliable and accurate results can be obtained. Though it is costly method, large areas can be covered. Personal bias of enumerators may influence the quality of the data.

(v) **By getting informations from correspondents :** In this method data are collected not by the investigators but by the local correspondents or agents, who will report the informations, collected by them to the investigator or collecting authority. When approximate results are desired instead of most accurate results, this method will be suitable because of economy and expediency. Newspaper agencies and various departments of Government adopted this method when regular informations are required from wide area.

1.4.2 Method of collection of secondary data

Secondary data are used by the investigators for their purpose collecting from office records, publications, reports, books, journals etc. These data may also be available in unpublished or manuscript from. As secondary data are not always reliable as primary data, they should not be used blindly but can be used with great caution and care. The following points are to be examined carefully before using these data :—

(i) object and scope of original enquiry,

(ii) the definition of units in which the data are collected,

(iii) sources of compiler's information and the method of collection,

(iv) the degree of accuracy achieved by the compiler,

(v) homogeneity and uniformity of data,

(vi) the reputation of the original investigator of the data.

If the data are found satisfactory in every respect, then only these data can be used safely for the required purpose.

1.4.3 Procedures for collecting data

There are two known procedures for collection of data :—

(1) **By complete enumeration** (or census survey) : When data are collected in respect of every individual or item of a population, i.e., a group of observations to a phenomenon under statistical investigation, the investigation is carried out by means of complete enumeration or census survey. During census operation, population of the country is enumerated and every individual is included in the operation.

(2) **By sample survey :** Because of limitation of time, money and labour data are collected through samples, i.e., a part of the population for an investigation and then the investigation is carried by means of sample survey. In auditing procedure the popular test check method is used to check a small number of entries taken from different parts of the book of accounts. This is an example of sample survey method.

Generally sample survey is more preferable than complete enumeration due to following reasons :—

(a) reduced cost in terms of money or in terms of man hours,

(b) greater speed of collection in sampling than in complete enumeration,

(c) greater coverage to collect more informations and to include more areas, due to use of highly trained personnel and specialized equipments,

(d) greater accuracy and reliability due to supervision or using better equipments,

(e) greater applicability in case of infinite or hypothetical population or when the quality of an item can be determined only by destroying the item in the process as in testing life of an electric bulb.

1.5 Classification of data

The term 'characteristic' means a quality possessed by an individual viz. height, weight, age, etc. and 'parameter' is a statistical measure of some characteristic of population and is obtained from population observations viz,

mean, variance of the population are the parameters. A statistic is a statistical measure of some characteristic of sample and is obtained from sample observations viz. mean and variance of the sample drawn from population are the statistics.

Classification of data is the process of arranging the data in groups or classes according to their to their resemblances or similarities. In the classification of data, units having common characteristic are placed in one class and thus the whole set of data is divided into a number of classes. The purpose of classification are (i) to facilitate meaningful comparison, (ii) to condense the data, (iii) to study the relationships.

Data can be classified as follows :—

(a) **On qualitative basis :** Qualitative characteristic of data is called attribute. This classification is done by arranging the data according to quality of it. Religion of a person is an attribute and then data are classified according to different religions viz. Hindu, Muslim, Christian, Jain, Buddhist etc. Here characteristic is non measurable.

(b) **On quantitative basis :** Quantitative characteristic of data is called variable. This classification is done by arranging the data according to numerical magnitude of data. So here characteristic is measurable. Height, weight of a person are the variables.

(c) **On geographical basis :** These data are arranged according to geographical region. These are also called spatial series data. Example of this is population of different districts of West Bengal.

(d) **On chronological bases :** These data are arranged according to order of time. These are also called Time series data. For example, production of paddy 5 consecutive years starting from 1991 in West Bengal is time series data.

If the variable takes some isolated values, finite or countable infinite in number then it is called discrete variable. For example, family size, number of road accidents in a road crossing in a month are discrete variables.

If the variable takes any value in an interval (i.e. uncountable infinite values in number) then it is called continuous variable. For example, height of a person, score of a student are continuous variables.

Frequency of a variable or an attribute means the number of times the values of the variable occur individually or in groups or how many times different qualities of an attribute, as the case may be, occurred in certain statistical investigation. Here the data are frequency data. It will be seen that most of our discussion will be devoted to the treatment of frequency data. Here identity of the individuals is unimportant, but we are interested in the characteristics of the groups formed by the individuals rather than those of individuals. For example, sex of 100 persons working in an office in 2004, number of seeds in 200 peapods, scores of 1st year students in mathematics of Annual examination of a college in 2004 are the frequency data.

If the identity of the individual values has to be kept in view in statistical study, then the data are called non-frequency data. For example, time-series data (or chronological data) and spatial series data (or geographically classified data) are non-frequency data.

1.6 Representation of data

Data may be presented in the following forms :—

1.6.1 Textual representation

This is a presentation of data in descriptive form i.e., in the form of presentation of words in a para of a Text book. As it is lengthy presentation, data cannot be compared at a glance and comprehending of main points is difficult.

1.6.2 Tabular representation

This is a systematic presentation of numerical data arranging vertically in columns and arranging horizontally in rows in accordance with some salient features. This representation facilitates comparison and analysis of data.

It helps to understand complex numerical data and makes the data simple and clear to separate their similar and dissimilar facts. It condenses data in the form of a table so that it may be easily understood and any comparisons involved there may be more readily made. It is a useful tool of analysis. Two or more tables may be necessary for presenting the data in simple and concise form.

A table has several parts which are described below :—

1. **Table number :** An appropriate number should be given to a table for its identification and easy reference in future.

2. **Title :** It is brief statement of the contents of the table and it should be placed at the head of the table :

3. **Stub :** This is extreme left part of the table. It gives the descriptions of rows of the table.

4. **Caption :** This is upper part of the table. It gives a description of various columns and sub-columns including their units of measurements.

5. **Body :** It contains the numerical information. It is most important part of the table. The arrangement of the body is done generally from left to right in rows and from top to bottom in columns.

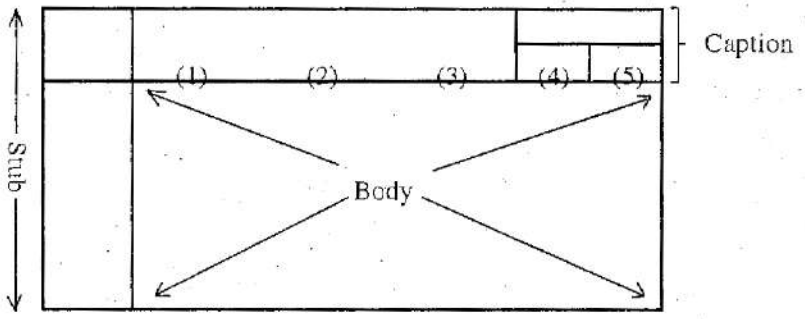
6. **Foot note :** This component clarifies some specific items of the table and is placed at the bottom of the table.

7. **Source :** This indicates the origin of the available information.

Sketch of a table showing its different parts :

Table No. .....

Title .....



Foot note (if any) .....

Source .....

Note : The Title, stub and caption taken together form the box head of the table.

1.6.2.1 Difference between classification and tabulation :

By both the process, classification and tabulation, the collected data are summarised and put in systematic order. For statistical investigation both are important. Classification is done first. Then they are presented in the tables. Classification is the basis for tabulation. In tabulation classified data are placed in rows and columns. Classification is a process of statistical analysis while tabulation is a process of presenting the data in suitable form.

1.6.2.2 Rule for tabulation :

Although there is no hard and fast rule to construct a table, common sense and experience should be employed. A good table must possess the following properties :

A table must contain an appropriate title. It must be simple and compact. It should not be overloaded with detail. If details are necessary, they should be shown in separate small tables mentioning their numbers. It should be well balanced in length and breadth. It should neither be too long vertically not be too short horizontally. The data should be arranged in systematic and logical manner. To facilitate comparisons, the figures to be compared should be placed close to each other either in columns or in rows as for as possible. The units of measurements, if any, for the items must be clearly mentioned in rows and columns. Light rulings should be used to separate sub-columns while heavy rulings should be used to distinguish main columns. Abbreviations should be avoided and ditlo marks should not be used. A miscellaneous column should be added to include unimportant items. Zero values or nonexisting values should be written as dash (—).

Example : Present the following information in tabular form :

"In 1965, out of a total of 2000 workers of a factory, 1500 workers were members of a trade union. The number of women employed was 150 of which 128 do not belong to the trade union. In 1970 the number of union workers increased to 1620 of which 1582 were men. On the other hand, the number of non-union workers fell down to 448 of which 318 were men."

Solution : Table No. 1

Table : No. of workers with sex and membership in trade union in 1965 and 1970.

Year	1965			1970		
Type of Workers \ Sex	Male	Female	Total	Male	Female	Total
Member of trade union	1478	22	1500	1582	38	1620
Not member of trade union	372	128	500	318	130	448
Total	1850	150	2000	1900	168	2068

Source : H.S. question paper of 1985 of H.S. Council, W. B. The informations are not given in the cells in which dashes are given and information in all other cells are given. The observations in the cells containing dashes are obtained as follows :

In 1965 total of workers who are not members of trade union = 2000 – 1500 = 500, No. of male workers = 2000 – 150 = 1850,

No. of male workers who are not member of trade union = 500 – 128 = 372

No. of male workers who are members of trade union = 1850 – 372 = 1478

No. of female workers who are members of trade union = 150 – 128 = 22

In 1970 total of all the workers = 1620 + 448 = 2068,

total of male workers = 1582 + 318 = 1900, Total of female workers = 2068 – 1900 = 168,

No. of female workers who are member of trade union = 1620 – 1582 = 38

No. of female workers who are not members of trade union = 168 – 38 = 130.

1.6.3 Diagrammatic representation

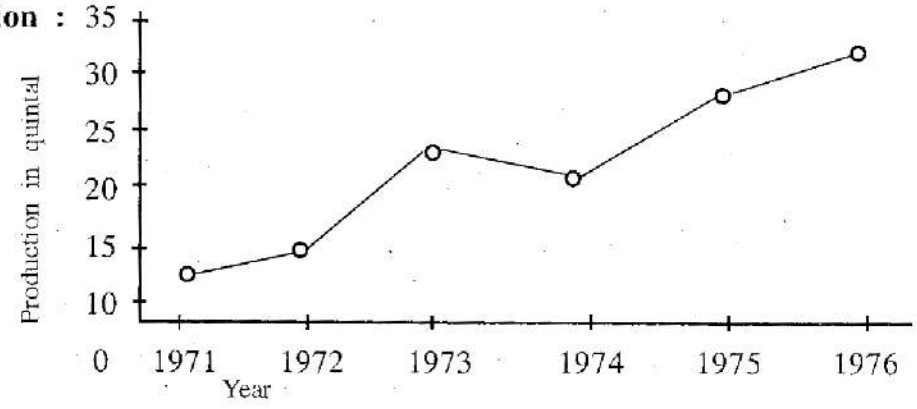
Though the classification and tabulation reduce the complexity of statistical data, still the data are not easily understood by the common people. If the mass of data is plotted graphically they become easy to understand to the common people. Also to get a vivid idea about the information given in a table, at a glance, the diagrammatic representation of the data is a better device. However charts or diagrams cannot represent a table in detail and can give only approximate positions. To represent a data set by diagrams takes much time than to represent it in tabular form. Some common types of diagrams generally used are—

(a) **Line diagram :** If the statistical data are given in accordance with the time of occurrence, line diagram is an appropriate representation of data. In plane paper or graph paper two perpendicular lines are drawn crossing at origin O. Horizontal line is taken as X-axis and time is taken along x-axis choosing suitable scale there. Vertical line is taken as Y-axis and the data viz. production, population etc. are taken along y-axis. There also suitable scale is chosen. Then the given data are represented by a set of points on the plane or graph paper. They are joined by straight lines to get line diagram of the data.

Example : Represent the following statistical information of a sugar factory by line diagram.

Year	1971	1972	1973	1974	1975	1976
Total production (in lakh quintal)	12	15	22	20	28	32

Solution :



Line chart of production of sugar in a sugar factory during the period 1971 to 1976. First take horizontal axis for the years and given years are considered at equal space. Then draw perpendicular vertical line for the production of sugar in quintals and we take some scale as marked in vertical line. Then plot points and encircled. Then the points are added by straight lines.

(b) **Bar diagram :** Bar diagram is composed of a number of equispaced bars of equal width, (suitably chosen), each for one category of given statistical data. Bars are drawn on the base line which may be a horizontal line or a vertical line. Lengths of rectangles or bars indicate the corresponding values of the statistical data and can be drawn considering suitable scale. Vertical bar diagram and horizontal bar diagram are named according as the base line is horizontal and vertical respectively. Generally, horizontal bar diagram is drawn in case of spatial series data and vertical bar diagram is drawn in case of time series data.

Bar diagrams are of following types :—

(i) **Simple bar diagram :** It consists of a number of equidistant vertical (or horizontal) bars of uniform width starting from horizontal axis (or vertical axis) and are shaded. A suitable scale is chosen and indicate it along vertical (or horizontal) axis. Length or height of the bars are considered according to the magnitude of given values. Then bars are drawn and shaded. Bars are generally vertical and horizontal if the data given are of a time series and of a spatial series respectively.

(ii) **Multiple bar** (or compound bar) **diagram :** It is a particular type of bar diagram, used to compare two or move series of related data such as male and female populations in five consecutive years. Here a number of sets of bars of equal width are drawn so that bars for a period or a related phenomenon are put together and uniform space is maintained between two or more sets of bars.

(iii) **Divided** (or component) **bar diagram :** When the data are available for a number of components in two or more situations (viz, two or more periods) and comparison among the different categories and the relation between each part and the whole is necessary in the diagram then this diagram is drawn. Here a single bar of appropriate length and width is drawn and its area is

considered as 100. The area of the bar is then divided into a number of parts depending on number of categories with the help of straight lines, drawn parallel to the base so that the area of the part indicates the percentage of the category concerned. Similarly other bars are drawn and the portions of the diagram are shaded differently for different categories. It shows relative sizes of different components of a whole in two or more situations.

Example : The following data show the profits of a factory from 1985 to 1990. Draw a suitable diagram.

Year :	1985	1986	1987	1988	1989	1990
Profit (in lakh Rs) :	15	18	20	15	13	17

Solution : The data is represented by vertical diagram profitability of factory

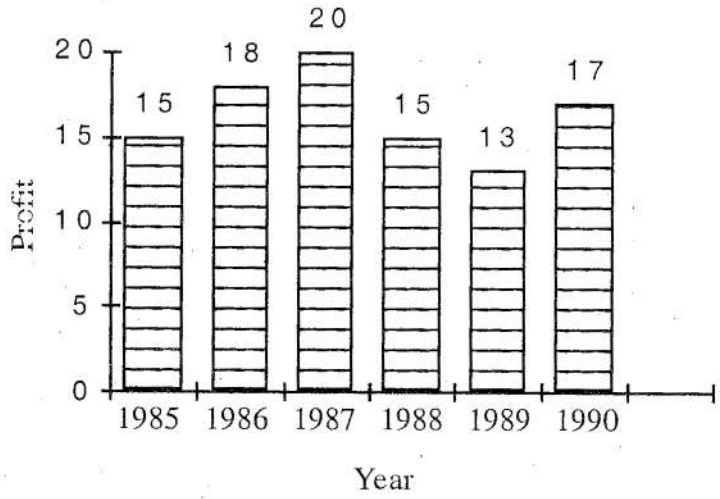


Fig : Bar diagram of profit of the given factory

Example : The following table gives the country of origin of feature films exhibited in India.

Country :	India	USA	UK	Other countries
No. of films :	144	81	64	16

Represent them by suitable diagram

Solution : The data is represented by horizontal bar diagram.

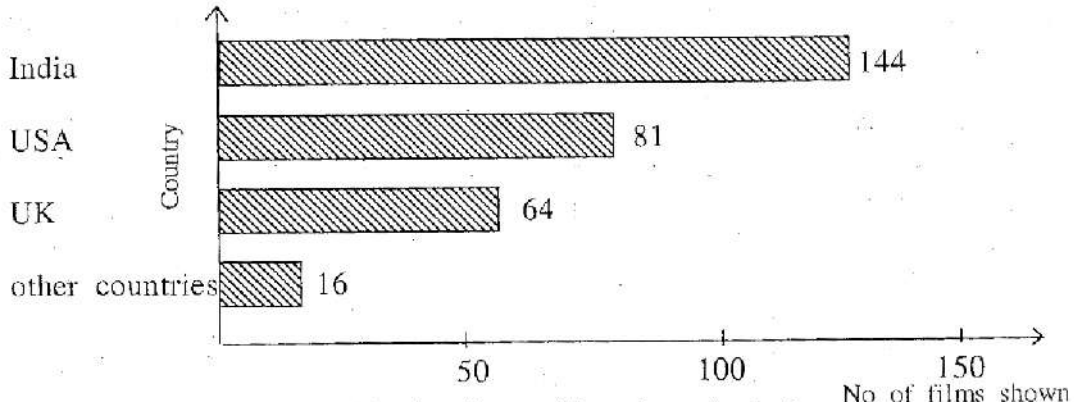


Figure : Showing feature films shown in India

Example : Make a graphical comparison of the compound bar diagram of the production of paddy corps (in lakh tons) for two different states A and B for the years 1981, 1991, 2001 from the following informations

Production of paddy in lakh tons—

Year ↓	States→ A	B
1981	40	35
1991	55	47·5
2001	30	50

Solution : We represent years along horizontal axis and production of paddy along vertical axis. Two states A and B form a block with two consecutive rectangles of bars with different shades. We maintain equal gap between composite blocks and all the blocks are of equal width.

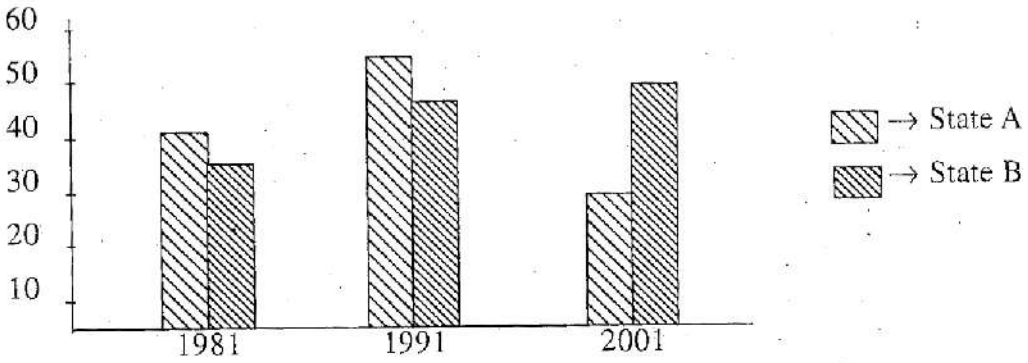


Figure showing production of paddy in two states A and B in 3 years

Example : Draw a two dimensional diagram to represent the following data.

Item of expenditure	Daily expenditure in Rupees.	
	Family 1	Family 2
1. Food	200	300
2. Clothing	50	72
3. Eduction	30	48
4. House rent	40	72
5. Miscellaneous	80	108
Total	**400**	**600**

Solution : The total expenditure will be taken as 100 and the expenditure on each item will be expressed in percentage. The width of the two rectangles will be in proportional to the total expenditure of the two families, i.e., 400 : 600, i.e., 2 : 3. The height of each rectangle will be same as it represents 100 percent.

Item of expenditure	Daily Expenditure in Rupees					
	Family 1			Family 2		
	Rs.	%	Cumulative %	Rs.	%	Cumulative %
A. Food	200	50	50	300	50	50
B. Clothing	50	12·5	62·5	72	12	62
C. Education	30	7·5	70	48	8	70
D. House rent	40	10	80	72	12	82
E. Miscellaneous	80	20	100	108	18	100
Total	**400**	**100**		**600**	**100**	

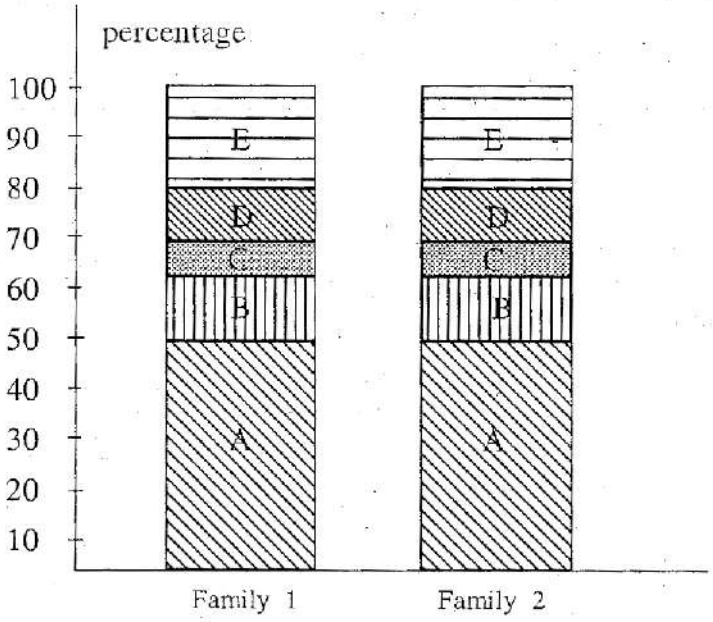


Figure showing expenditures on different items with totals in two families

(c) **Pie diagram :** It represent the component parts of an aggregate. The diagram gives the comparison between the various components between a part and the whole. In fact a circle is drawn at first and then it is subdivided into different sectors by calculating percentages of different components considering whole as 100%. Practically a pie diagram is drawn by calculating degree for the components in the centre of the circle when whole circle form 360° in the centre i.e., by multiplying percentages of each component and the whole by 3·6. First angle of one component is drawn at the centre of the circle by a divider.

Then another angle is drawn at the centre adjacent to the former angle and above it by setting divider on upper side of the angle for another component. Thus all the angles are drawn to make angle 360° at the centre and the different sectors are shaded differently to form a pie diagram. Angles can also be determined for each component by multiplying 360° to the relative amount of each with respect to total.

Example : Represent the following data by a pie diagram.

Item	Expenditure (in Rs.)	Item	Expenditure (in Rs.)
Food	87	Education	13
Clothing	24	House rent	25
Recreation	11	Miscellaneous	20

Solution :

Item	Expenditure	Angle of the circle
A) Food	87	$\frac{87}{180} \times 360° = 174°$
B) Clothing	24	$\frac{24}{180} \times 360° = 48°$
C) Recreation	11	$\frac{11}{180} \times 360° = 22°$
D) Eduction	13	$\frac{13}{180} \times 360° = 26°$
E) House rent	25	$\frac{25}{180} \times 360° = 50°$
F) Miscellaneous	20	$\frac{20}{180} \times 360° = 40°$
Total	**180**	**360°**

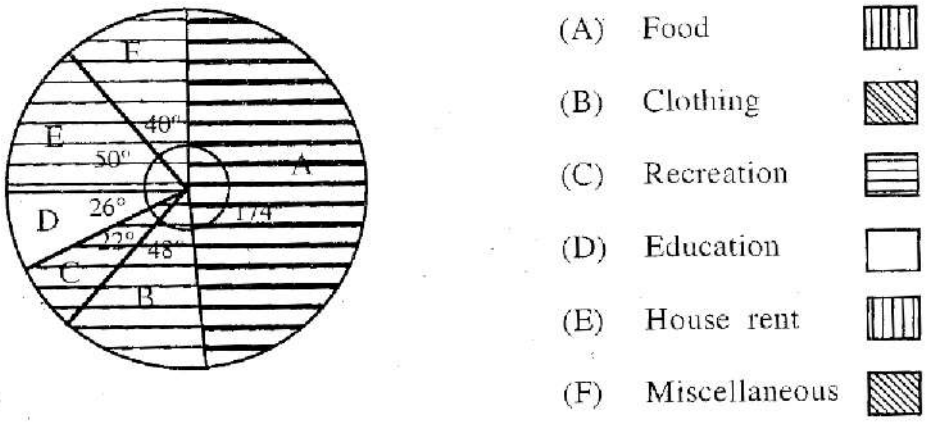


Figure of Pie diagram showing expenditure of different items along with its total

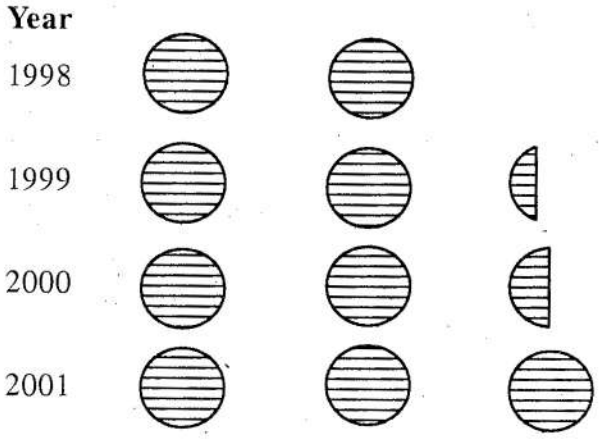
(d) **Pictogram and Cartogram :** Pictogram is a device of representing data is pictures. This diagram is popular and easily understood. They are extensively used by the Government and private organisations to compare the data of different categories.

Cartogram is used to present the statistical data through maps accompanied by various types of diagramatic representation. Cartogram is simple and easy to understood. These are generally used when the geographic comparisons are to be made.

Example : Draw a pictogram of the number of tourists visiting a place in thousands.

Year	:	1998	1999	2000	2001
No. of tourists	:	16	18	20*	24

Solution : To draw the pictogram of the above data, assume that a complete circle represents 8000 tourists. So, the pictogram will be as follows.



(e) **Ratio chart :** It is a kind of line diagram where vertical scale is logarithmic and horizontal scale is of natural or arithmetic type. This chart is also called semi-logarithmic chart. So horizontal scale give absolute changes of values, but vertical scale give relative changes of values. As values are in geometric progression then consecutive terms in order are in common ratio r. Then logarithm of the corresponding terms in order are in common difference which is log r. Thus if the variable under enquiry is increasing or decreasing

at a constant ratio then the ratio chart will be exactly a straight line. This chart is generally used in time series data of population where the data are changing with time in constant ratio. This chart is obtained (i) by plotting the values of x (i.e., abscision) and y (i.e., ordinate) on a special type of graph paper in which the natural scale is used in horizontal direction and the ratio or logarithmic scale is used in vertical direction or (ii) by plotting the value of x and log y in ordinary graph paper using natural scale in both directions.

Example : Represent graphically from the following data the growth of the population of a particular state to show both relative growth

Ceusus Year	:	1901	1911	1921	1931	1941	1951
Population (in lakhs)	:	50·0	52·4	55·6	59·6	65·0	68·7

Solution :

Calculation of logarithms

Year	y = Absolute value of population (in lakhs)	Logarithm (= log y)
1901	50·0	6·6990 = 6·70 (approx)
1911	52·4	6·7193 = 6·72 ,,
1921	55·6	6·7451 = 6·75 ,,
1931	59·6	6·7745 = 6·77 ,,
1941	65·0	6·8129 = 6·81 ,,
1951	68·7	6·8370 = 6·84 ,,

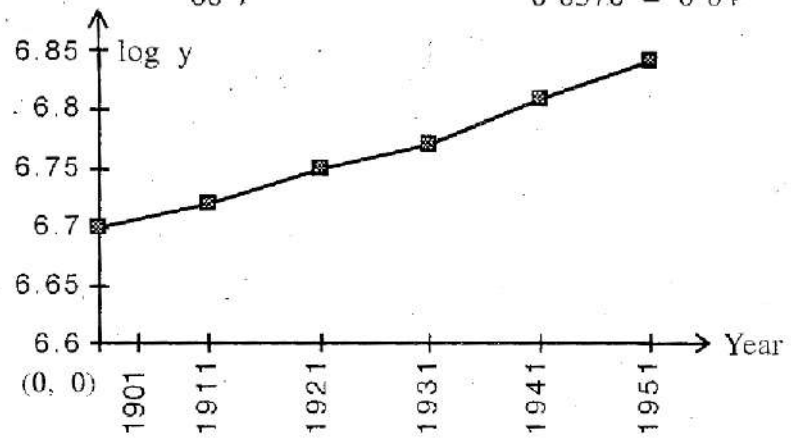


Figure showing ratio chart giving relative growth of population

1.7 Summary

This chapter consists of definitions of statistics and its applications, limitations, collection of data including their methods, classification of data, different representations of data. Different examples are given for clear understanding of the contents of this chapter.

1.8 Exercises

1. What do you mean by statistics? What are data?
2. Write the different stages of statistical investigation of data.
3. Write the different functions of statistics.
4. In what fields statistics can be used? Write them clearly.
5. What are the limitations of statistics?
6. Define primary data and secondary data. How they can be collected?
7. How many representation of data do you know? Give their name.
8. Distinguish between (i) sample survey and complete enumeration, (ii) variable and attribute.
9. What is meant by tabulation? What are different parts of a table?
10. Compare classification and tabulation.
11. What are the general rules to construct a table?
12. Describe different methods of representation of data and compare between them.
13. What is diagramatic representation? Discuss its advantages.
14. What do you mean by a Bar Chart? How is it drawn?
15. What are line charts and ratio charts? Distinguish between them.
16. Write short notes on
    (i) Line diagram.
    (ii) Bar diagram.
    (iii) Pie diagram.

(iv) Pictogram.

(v) Ratio chart.

17. What is multiple bar chart?

18. Define Horizontal Bar chart and Vertical Bar chart. When are they used?

19. Define component bar diagram and pie diagram and compare between them.

20. Describe different types of bar diagram. Describe the situations and techniques of their applications.

21. Draw a blank table of exports and imports during 5 years 1990—1994 relating to ports at Bombay, Calcutta, Madras and other ports.

22. Draw up a blank table to show the values of exports from India to U.K., U.S.A., U.S.S.R., Canada and West Germany over the five consecutive years 1971 to 1975.

23. Draw a blank table to show the number of students admitted in classes 1st year and third year in each of three streams Arts, Science and Commerce in two educational years of a college of male students showing totals in each class, stream and educational year.

24. Draft a blank table to show the distribution of male and female employees of age groups —below 25 years, 25-40 years and 40-60 years working in an office for the last three years in the income groups below Rs. 1000, 1001—2550, 2501—5500 and above 5000.

25. In a sample study about tea habits of people in two towns A and B the following data were observed. Tabulate them.

Town A	**Town B**
50% were males	55% were males
80% were tea drinkers	72% were tea drinkers
45% were male tea drinkers	38% were male tea drinkers

26. Number of men workers, women workers and total workers in a factory are given in the years 1970, 1975 and 1980 separately for members and non-members of a trade union. Draw a blank table to exhibit the data.

27. "1652 men and 1226 women participated in a poll on the opinion about a certain measure. 1006 persons of whom 796 were male, voted against the measure. In all 1425 persons voted for the measure, 256 women were indifferent." Tabulate these data in a suitable tabular form and find what percentage of men were for the measure.

28. Represent the following statistical information graphically.

Year :	1934	1935	1936	1937	1938	1939	1940	1941
Monthly average production :	619	532	215	618	561	642	526	502

29. Draw a bar diagram to represent the following data.

Year	1991	1992	1993	1994	1995	1996
Values of imports in crores Rs.	2350	2319	2346	2433	2561	2612

30. Draw a bar chart for the number of students of a college :

Higher secondary commerce class	360
B. com 1st year class	610
B. com 2nd year class	660
B. com 3rd year class	310

31. Draw a suitable diagram from the following data.

Year	Sale ('000 Rs.)	Gross Profit ('000 Rs.)	Net Profit ('000 Rs.)
1994	235	45	25
1995	245	55	30
1996	260	65	35
1997	280	70	34
1998	305	80	40

32. Draw appropriate diagram for the following data :

Production of tea in different countries

Country	**Production ('000 metric tons)**
India	425
Sri Lanka	245
Japan	90
Indonesia	55
Other countries	385
Total	**1200**

33. The following table gives the total cost in rupees and its component parts in two periods for the same item.

Item	Cost in 1980 (in Rs)	Cost in 1990 (in Rs)
Direct Raw Material	50,000	60,000
Direct Labour charge	60,00	70,000
Machining Expense	20,000	20,000
Overhead expense	35,000	50,000
Total	**1,65,000**	**2,00,000**

34. Draw a pictogram of the following data of number of tourists visiting a place given—

Year	No. of tourists ('000)
1988	16
1989	18
1990	20
1991	24
1992	28

35. Plot the following data relating to population of India so as to indicate the proportionate increase in population from one period to another :

Year :	1871	1881	1891	1901	1911	1921	1931
Population (in millions) :	200	250	310	390	490	610	763

1.9 Suggested Reading

1. Chaudhuri, S. B. Elementary Statistics, Vol I, Shraddha Prakashani 1986.
2. Croxton, F. E. and Cowden, D. J. Applied General Statistics, Prentice Hall, 1964.
3. Gun, A. M; Gupta, M. K. and Dasgupta, B. Fundamentals of Statistics, Vol.-I, World Press Pvt. Ltd. 2002.
4. Mukherjee, A. Fundamentals Treatise on Probability And Statistics. Sree Tara Prakashani, 2003.
5. Fisher, R. A statistical methods for Research workers, Oliver & Boyd, 1954.

Unit 2 ❑ Frequency Distribution

Structure

2.0 Objectives

After collection, data are arranged in arbitrary manner which we call raw data. This type of presentation does not help workers in the line of statistics directly. The data then should be classified and arranged in suitable form so as to read and understand them easily. One suitable form of this type is frequency distribution of an attribute or a variable. These distributions are also required

to diagramatise in graph or plain paper for easy and quick appreciation of them and for different statistical analysis.

2.1 Introduction

Organising and summarising the data are two important functions of statistics. Here we give necessary rules and guidelines for constructing a frequency distribution and for drawing their graphs.

Frequency distribution represents a definite form of representing and summarising the data. It serves the basis for developing statistical methods. Graphical representation of frequency distributions are considered on the basis of observations classified and tabulated according to their size and magnitude. First data are collected through a sample survey or complete enumeration. The data are then arranged in an organised and summarised form in such a way that they become easy to read, understand, assimilate and to highlight the basic trends and broad variations. Frequency distribution is one such form.

A frequency distribution gives better grasp over the data and facilitates more efficient data analysis to make proper decision making.

2.2 Frequency distribution

Qualitative characteristic of data is called attribute and quantitative characteristics of data is called variable. The variables are of two types, discrete variable and continuous variable. A variable is called discrete when it takes finite or countably infinite number of values and it is called continuous when it is capable of taking any value within a specified interval i.e., uncountable infinite number of values in that interval. For example, sex of a person, smoking habit of a person are attributes, number of births in different years, number of telephone calls in different hours of a day are discrete variables and score of a student, height of a person are continuous variables. In case of continuous variable discreteness considered is completely artificial due to limitation of measuring instruments.

Data arising in different classes of an attribute are data of attribute. Data which are the values of a discrete variable are called discrete data and data which describe the continuous variable are called continuous data. In general measurements generate continuous data and enumerations or counting generate discrete data. The number of times class of an attributes occurs in a statistical investigation is called the frequency of the class of that attribute. Similarly, the number of times the values of a variable, discrete or continuous, occur in an in statistical investigation is called frequency of those values. Generally continuous variables are considered in groups. Then number of observations falling in a group is called frequency of that group or class. Observations or observed values are actually the values of a variable obtained by observation in an investigation. Then we construct a frequency distribution of an attribute or a variable.

The statistical data recorded in an arbitrary manner from a field of enquiry is also called raw data. The raw data requires classification showing different classes with their frequencies in case of an attribute and classification of different values or groups of values with their respective frequencies in case a variable. This classification with respective frequencies gives frequency distribution. It is frequency distribution of an attribute if the classification is done for an attribute and of a discrete variable or continuous variable if the classification is done in terms of values or groups of values of the discrete or continuous variable.

To know the smoking habit of people in a village A of West Bengal each individual of that village has been asked whether he or she is a smoker or not, the data are then arranged in the frequency distribution table of attribute "Smoking habit of 1100 people residing in village A of West Bengal" from the raw data of observations from the village like "S, N, S, S, S, N, N, S, S, S, N" by using tally marks in group of five where S = smoker, N = non-smoker.

First from the table below with classes then consider on tally work for class smoker from observation S and one tally mark for class non-smoker from observation N and the procedure is repeated to form groups of tally marks each group containing 5 tally marks.

Table-1

Results of smoking habit in village A

Smoking habit	**Tally mark**	**No. of people**
Smoker (S)	𝍸 𝍸 ... 𝍸 |||	423
Non-Smoker (N)	𝍸 𝍸 𝍸 ... 𝍸 ||	677
Total		**1100**

So frequency distribution of smoking habit in village A is the following distribution table :

Smoking habit	No. of people
Smoker	423
Non-Smoker	677
Total	**1100**

A frequency distribution of a discrete variable is obtained by using tally marks from a set of raw data of discrete values i.e., number of peas (the discrete variable) in 53 peapods :

2, 1, 3, 6, 5, 4, 5, 3, 2, 4, 5, 4, 3, 5, 4, 3, 3, 5, 3, 2, 5, 3, 4, 4, 4, 3, 3, 5, 2, 2, 3, 3, 4, 5, 2, 5, 6, 2, 1, 2, 3, 3, 4, 4, 3, 3, 2, 3, 4, 3, 3, 4, 2.

To form the frequency distribution of number of peas we start as follows. We consider one raw data and give tally mark corresponding to the value of variable. Make tally marks a group of five.

No. of peas	Tally mark	Frequency (No. of peapods)
1	||	2
2	𝍸 𝍸	10
3	𝍸 𝍸 𝍸 |||	18
4	𝍸 𝍸 ||	12
5	𝍸 𝍸	9
6	||	2
Total		**53**

So the frequency distribution of number of peas is—

No. of peas	:	1	2	3	4	5	6	Total
No. of peapods	:	2	10	18	12	9	2	53

As there are no group of values of variable, it is called simple frequency distribution. Here values of the variable are written individually with their frequencies. In case of grouped frequency distribution values of the variable are arranged in groups with corresponding frequencies. Otherwise table will be elongated with different values of the variable. Then the data are condensed by putting the values of the variable into small number, say, between 6 and 15, of groups or classes of same or different sizes, considered suitably from the observed values of the variable. The number of values belonging to each groups (also called classes or class intervals) is called class frequency and they can be obtained by using tally marks in groups of five. This frequency distribution is called grouped frequency distribution of the variable. The groups are called class interval or class.

Class limits : The class limits of a class of values are the smallest and largest values of the class considered. For example, take the class interval 10—20 the lower class limit is 10 and upper class limit is 20.

Class boundaries : It the upper class limit of any class is not equal to lower class limit of the next class then to get any continuous graphical representation of data it is required to rearrange the class limits in such a way that the upper class limit of any class coincides with lower class limit of the successive class. Then these values which are averages of upper class limit of a class and lower class limit of next class are called class boundaries. Upper value of that class formed is called upper class boundary of that class and lower value is called lower class boundary of that class. In case of starting class the lower boundary will be obtained by subtracting half of the difference of lower limit of 2nd class and upper limit of starting class from the lower limit of the first class. In case of last class of the frequency distribution upper limit will be obtained by adding half of the difference of lower limit of last class and upper limit of former class.

Class marks : They are the mid-values of the class i.e., average of two class limits of the class.

Class width or length : Class width or length of a class interval is defined to be the difference between the lower and upper class boundaries (not the class limits) of that particular class interval.

Generally to prepare a grouped frequency table, the class width of each class is taken to be the same for simple computations. But when some of the observed values are few and far away from the rest, classes of unequal width may be considered. The following table shows how class boundaries, class marks class width can be determined.

Table

Class limits	Class boundaries	Class marks or mid-value	Class length
10-19	9·5-19·5	14·5	10
20-29	19·5-29·5	24·5	10
30-39	29·5-39·5	34·5	10
40-49	39·5-49·5	44·5	10
50-59	49·5-59·5	54·5	10
60-69	59·5-69·5	64·5	10
70-79	69·5-79·5	74·5	10

Frequency of a class is defined as the number of values falling within that class. Sum of frequencies of all the classes considered is called total frequency.

Frequency density of a class is defined as its frequency per unit width and is given by the formula,

$$\text{Frequency density} = \frac{\text{class frequency}}{\text{class width}}$$

Relative frequency of a class is defined as the ratio of class frequency to the total frequency and is given by the formula,

$$\text{Relative frequency} = \frac{\text{class frequency}}{\text{Total frequency}}, \quad \text{Percentage frequency} = 100 \times \text{Relative frequency.}$$

So frequency densities and relative frequencies are calculated as follows.

Table

Class boundary	Frequency	Frequency density	Relative frequency
9·5-19·5	7	0·7	0·07
19·5-29·5	10	1·0	0·10
29·5-39·5	30	3·0	0·30
39·5-49·5	40	4·0	0·40
49·5-59·5	9	0·9	0·09
59·5-69·5	4	0·4	0·04
Total	**100**		**1·00**

2.3 Guidelines for preparing a frequency table

No. of classes should not be too large or too small. In practice number of classes should be between 5 and 15 depending on the number of given observations. Some authors advice to follow the sturges rule

$$n = 1 + 3{\cdot}32 \log N$$

where n = number of classes and N = total frequency.

First calculate the range = highest value – lowest value. Divide the range by number of classes to be chosen for this frequency table and get the width of each class approximately. Select the width of the class a round number greater than or equal to actual width Generally consider frequency distribution of equal width for easy compilation of statistical measures. Width of the classes are generally taken as multiple of 5 i.e., 5, 10, 15 etc. classes should be non-over lapping so that no value comes under move than one class. No value should escape classification. Take one observation and give tally mark in the class it belongs and the procedure is repeated making tally marks in a group of five. From the tally marks calculate frequencies of the classes by counting. Thus obtain the grouped frequency distribution of the variable. If the data are given in one decimal, then consider one decimal place for the limits of the classes.

Example : Using sturges rule $n = 1 + 3{\cdot}322 \log N$ classify in equal intervals the following data of hours spent in working by 50 workers for period

of a month in a certain factory.

110	175	161	157	155	108	164	128	114	178
165	133	195	151	71	94	87	42	30	62
130	156	167	124	164	146	116	149	104	141
103	204	162	149	79	113	69	121	93	143
140	144	187	184	197	87	40	122	203	148

Determine mid values, class boundaries, frequency density, relative frequency of the frequency distribution.

Solution : n = no. of classes, N = no. of observations = 50

so $n = 1 + 3{\cdot}3222 \log 50 = 1 + 3{\cdot}3222 \times 1{\cdot}6990$

$= 1 + 5{\cdot}6441 = 6{\cdot}6441 = 7$

$i = \frac{\text{Range}}{n} = \frac{204-30}{7} = 24{\cdot}85$ or 25.

Class	Tally Marks	Frequency
30-54	|||	3
55-79	||||	4
80-104	𝍸 |	6
105-129	𝍸 ||||	9
130-154	𝍸 𝍸 |	11
155-179	𝍸 𝍸 |	11
180-204	𝍸 |	6
Total		**50**

So frequency distribution of hours spent by workers of a factory in a month.

Class of hours spent	Frequency
30-54	3
55-79	4
80-104	6
105-129	9
130-154	11
155-179	11
180-204	6
Total	**50**

Here class width of a class = upper class boundary – lower class boundary = 25

Frequency density = $\frac{\text{frequency}}{\text{class width}}$, Relative frequency = $\frac{\text{frequency}}{\text{Total frequency}}$.

Class in terms of boundaries	Frequency	Mid-value	Frequency density	Relative frequency
29·5-54·5	3	42	0·12	0·06
54·5-79·5	4	67	0·16	0·08
79·5-104·5	6	92	0·24	0·12
104·5-129·5	9	117	0·36	0·18
129·5-154·5	11	142	0·44	0·22
154·5-179·5	11	167	0·44	0·22
179·5-204·5	6	192	0·24	0·12
Total	50			1

2.4 Cumulative frequency distribution

Cumulative frequencies are of two types, (a) cumulative frequency less than type (in short C. F. (< type)) and (b) cumulative frequency more than type (in short C.F. (> type)). The less than cumulative frequency for any value (class) is obtained by cumulating successively the frequencies of all the previous values (classes) including that value (class). The more than cumulative frequency for any value (class) is obtained by cumulating successively the frequencies of all the proceeding values (classes) including that value (class). Cumulative frequencies expressed as a percentage of the total frequency is known as cumulative percentage. The distribution of the variable including classes and cumulative frequencies are called cumulative frequency distribution. For the frequency distribution of just former example the cumulative frequency distributions less than type and more than type are given below.

Class (in boundaries)	Frequency	C.F. (< type)	C.F. (> type)
29·5-54·5	3	3	47 + 3 = 50
54·5-79·5	4	3 + 4 = 7	43 + 4 = 47
79·5-104·5	6	7 + 6 = 13	37 + 6 = 43
104·5-129·5	9	13 + 9 = 22	28 + 9 = 37
129·5-154·5	11	22 + 11 = 33	17 + 11 = 28
154·5-179·5	11	33 + 11 = 44	6 + 11 = 17
179·5-204·5	6	44 + 6 = 50	6

The table shows that the number of observations below :—

29·5, 54·5, 79·5, 104·5, 129·5, 154·5, 179·5, 204·5 are 0, 3, 7, 13, 22, 33, 44, 50 respectively and the number of observations above 204·5, 179·5, 154·5, 129·5, 104·5, 79·5, 54·5, 29·5 are 0, 6, 17, 28, 37, 43, 47, 50 respectively.

2.5 Graphical representation of frequency distribution

2.5.1 Frequency bar diagram or Column diagram

To represent the frequency distribution of a discrete variable graphically, one uses column diagram or a frequency bar diagram where the values of the discrete variable are placed on the horizontal axis and the corresponding frequencies are placed on the vertical axis considering suitable scales. Perpendiculars representing frequencies are drawn on the points representing the values of the variable on the horizontal axis.

2.5.2 Frequency polygon

Use of frequency polygon is another method of representing frequency distribution of a discrete variable. Considering two dimensional coordinate axis in a graph paper or an ordinary paper, points with values of discrete variable as abscissa and frequency of them as ordinate are first drawn on the graph

paper considering suitable scales in both cases. Joining those points by straight lines and joining first point and last point with the points on the horizontal axis at equal distances moving backward from 1st two given values and moving forward from last two given values of discrete variable.

Frequency polygon may also be used to represent the frequency distribution of a continuous variable provided the classes of the variable are of equal length. The frequencies are then plotted vertically against mid values of the corresponding classes of the continuous variable, plotted horizontally and the points thus obtained are joined by straight lines. Then first point and last point are joined to the two points on the horizontal axis having zero frequencies. Of these two points on the horizontal axis, one is the mid point of the just earlier assumed class of equal length as of the first class and the other is the mid point of just next assumed class of equal length as of the last given class.

2.5.3 Histogram

A histogram is used to describe the grouped frequency distribution of both continuous and discontinuous type diagramatically. Class boundaries of the groups are plotted in the horizontal axis considering suitable scale. Rectangles whose heights are represented by the class frequency densities are drawn vertically over the corresponding class intervals so that area of each rectangle would be the corresponding frequency of that class interval. If the class lengths of the given groups are equal then frequency densities are proportional to the frequency of the classes. Then frequencies can be considered vertically instead of frequency densities since both will give the same diagram. Histogram is a diagram of series of adjoining rectangles.

2.5.4 Cumulative frequency diagram or ogive

It is a diagramatic representation of a simple or grouped frequency distribution on the basis of cumulative frequencies. There are two types of ogives and they are less than type ogive and more than type ogive.

Consider the case of grouped frequency distribution. Then to draw the less than type ogive first plot the upper boundaries of the classes and the lower boundary of the first class along horizontal axis and plot the corresponding less than type cumulative frequencies along vertical axis to get the different points.

In this context it is need to be mentioned that cumulative frequency (less than type) for the lower class bouradary of the first class is zero when the values in the classes are in increasing order. All these points are joined by different line segments and two horizontal lines are drawn one perpendicular to the horizontal axis from the last point towards the right and another along horizontal axis from the point drawn on horizontal axis towards left. This diagram looks like elongated S.

Similarly to draw more than type ogive there points are drawn first with lower class boundary along the horizontal axis and more than type cumulative frequency along vertical axis. Consider atlast a point on the horizontal axis, whose abcissa is the lower class boundary of the last class and ordinate is zero. Join all the points plotted by line segments and draw one line perpendicular to the herizontal axis towards left from the first point plotted and a line along horizontal axis from the last point plotted. Then the diagram will be inverse elongated S type. This diagram is called more than type ogive.

If the above two ogives are drawn in the same graph paper, the abcissa of the intersection point will be the median of the corresponding variable. Ogives are used to determine different quartiles, deciles, percentiles and quantiles, cumulative frequency of any value of the variable and also the frequency between two given values of the variable.

Simple frequency distribution is generally considered for a discrete variable. Then less than type cumulative frequencies are plotted along vertical axis and the corresponding values of the variable (below which value the frequency is the said cummulative frequency) along horizontal axis to get points in a graph paper considering suitable scales in both the axes. The points are joined so as to form stair case (i.e., steps) ascending from left to right where height of the first step is the first less than type cumulative frequency, height of the second step is the second less than type cumulative frequency and so on. Here the values of the discrete variable are in increasing order. The variate value just below the lowest given value of the discrete variable with 0 less than type cumulative frequency is considered as a point on the horizontal axis. So first step starts from the horizontal axis. For maximum value of the variable and above draw a straight line parallel to the horizontal axis. This diagram is also called a step diagram.

In case of more than ogive, first form a more than type cumulative frequency distribution table so that above the largest value of the variable frequency is zero. Plotting points with values of the variable as abcissa and corresponding cumulative frequency (more than type) as ordinate points are joined so as to form stair case descending from left to right where height of the first step is the first more than type cumulative frequency, that of second step is the second more than type cumulative frequency and so on till the horizontal axis is reached in this fashion. Cumulative frequency (more than type) of highest value is zero and lines are drawn one parallel to horizonal axis from lowest of the values whose more than type cumulative frequency is the total frequency and other along horizontal axis from the highest value of the variable. The diagram thus obtained is called more than type ogive.

When more than type and less than type are drawn in same graph paper their intersection points give the median as abcissa. From any ogive quartiles, deciles, percentiles, and quantiles can be determined.

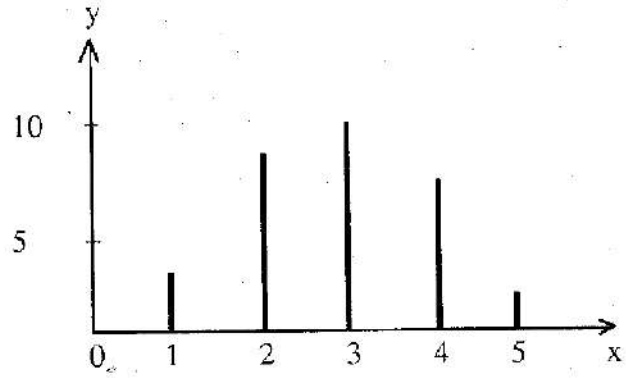
Example : From the following frequency distribution of daily number of car accidents represent the data by a suitable diagram. Also draw two ogives.

Table : Frequency distribution of no. of car accidents in 30 days.

No. of car accidents	No. of days
1	3
2	8
3	10
4	7
5	2
Total	30

Solution : The above data can be represented graphically either by frequency bar diagram or column diagram or by frequency polygon.

Column diagram : Groups of vertical thick bars of heights as frequencies i.e., the number of days at the corresponding points i.e., values of the variable 'number of accidents' on the horizontal axis. So suitable scales are considered on both the axis, horizontal and vertical.



x = no. accidents in a day and y = no. of days.

Fig. : Frequency bar diagram showing simple freqency distribution of daily no. of accidents on 30 days.

Frequency polygon : Considering x-axis and y-axis perpendicular to each other with suitable scale, plot the points (0, 0), (1, 3), (2, 8), (3, 10), (4, 7), (5, 2), (6, 0) and add them by line segments to get frequency polygon where x = no. of accidents in a day and y = no. of days.

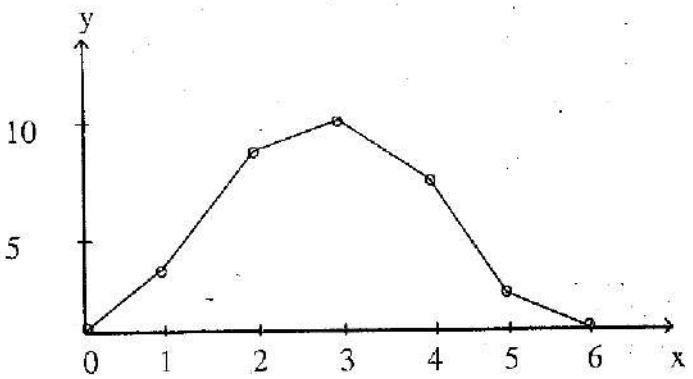


Fig : Frequency polygon showing the simple frequency distribution of daily no. of accidents on 30 days.

Ogives : First form a table of cumulative frequency distribution of both less than type and more than type.

Table : Cumulative frequency districtution of no. of car accidents.

No. of car accidents	frequency	Cumulative frequency	
		less than type	more than type
0	0	0	30
1	3	3	30
2	8	11	27
3	10	21	19
4	7	28	9
5	2	30	2
6	0	30	0

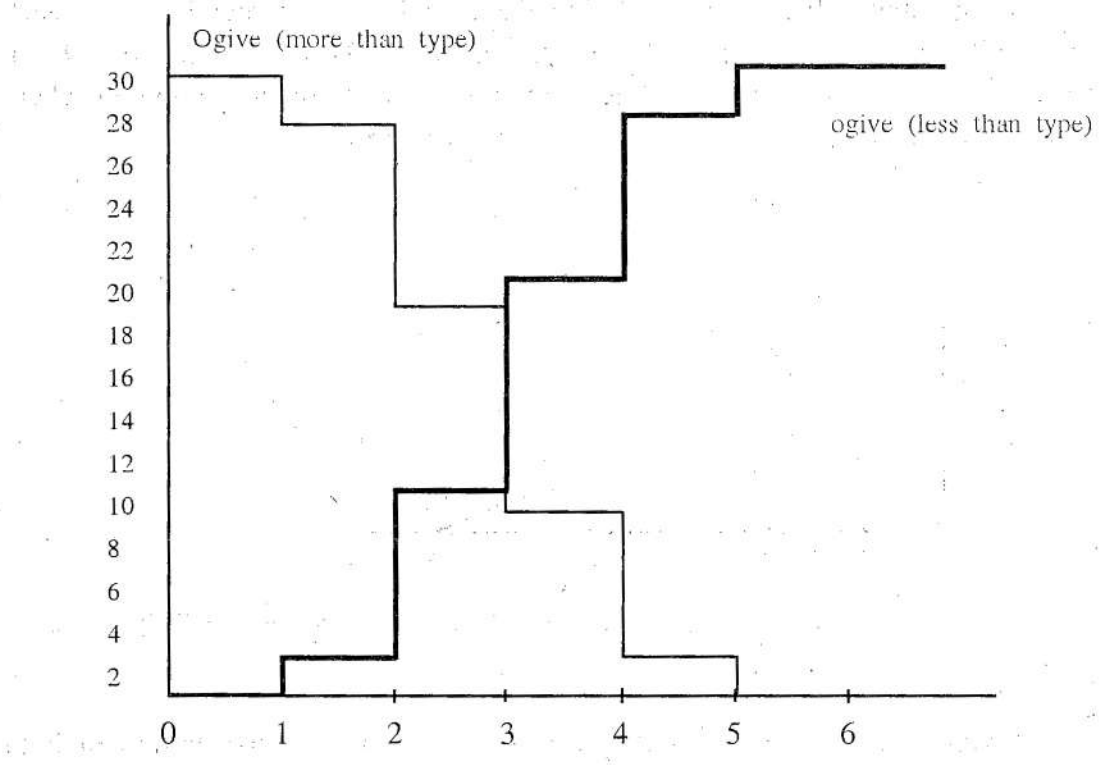


Fig. : Ogives for less than type (fat lined step diagram) and for more than type (thin lined step diagram).

Example : From the following distribution of wages of 100 workers of a factory in a day, draw the histogram, frequency polygon, less than type ogive and more than type.

Table : Wage distribution of 100 workers

Wage (in Rs.)	No. of workers (frequency)
100-109	10
110-119	15
120-129	30
130-139	25
140-149	15
150-159	5
Total	100

Solution : Now class lengths of the given classes are equal. We first obtain the classes in a frequency distribution in terms of class boundaries. Frequency densities here are proportional to the frequencies. Then class boundaries of wages are taken along horizontal axis and frequencies are taken along vertical axis. For frequency polygon mid points of the classes are required.

Wage-class	Mid point	Frequency
89·5-99·5	94·5	0
99·5-109·5	104·5	10
109·5-119·5	114·5	15
119·5-129·5	124·5	30
129·5-139·5	134·5	25
139·5-149·5	144·5	15
149·5-159·5	154·5	5
159·5-169·5	164·5	0
Total		100

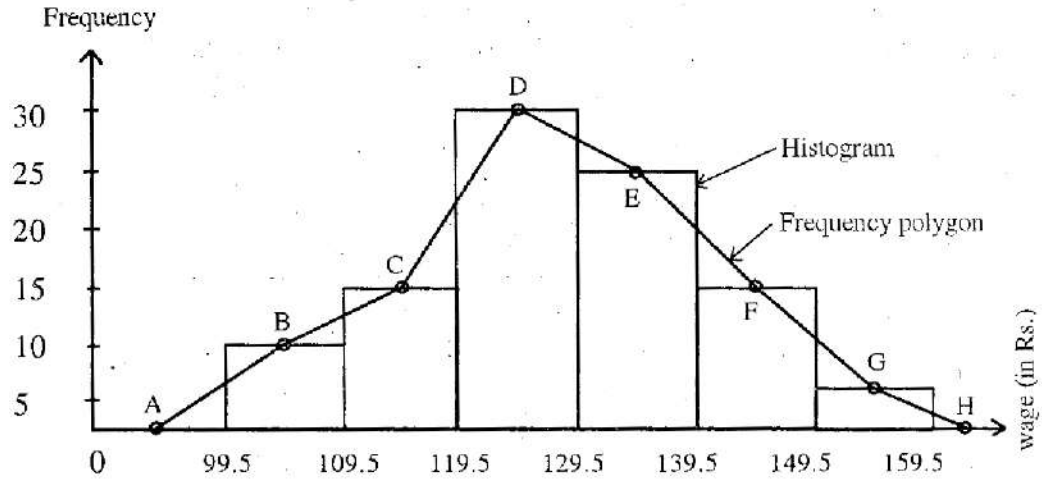


Fig. : Histogram and frequency polygon of wage distribution.

The whole area bounded by the adjacent rectangles represents histogram. The whole area dimarkated by the polygon is the frequency polygon which is obtained by joining the points A(94·5, 0), B(104·5, 10), C(114·5, 15), D(124·5, 30), E(134·5, 25), F(144·5, 15), G(154·5, 5), H(164·5, 0) by line segments.

Ogives : First obtain the cumulative frequency table both less than type and more than type.

Class of wages (class boundaries)	Freq.	Cumulative freq (C.F.)	
		less than type	more than type
89·5-99·5	0	0	100
99·5-109·5	10	10	100
109·5-119·5	15	25	90
119·5-129·5	30	55	75
129·5-139·5	25	80	45
139·5-149·5	15	95	20
149·5-159·5	5	100	5
159·5-169·5	0	100	0

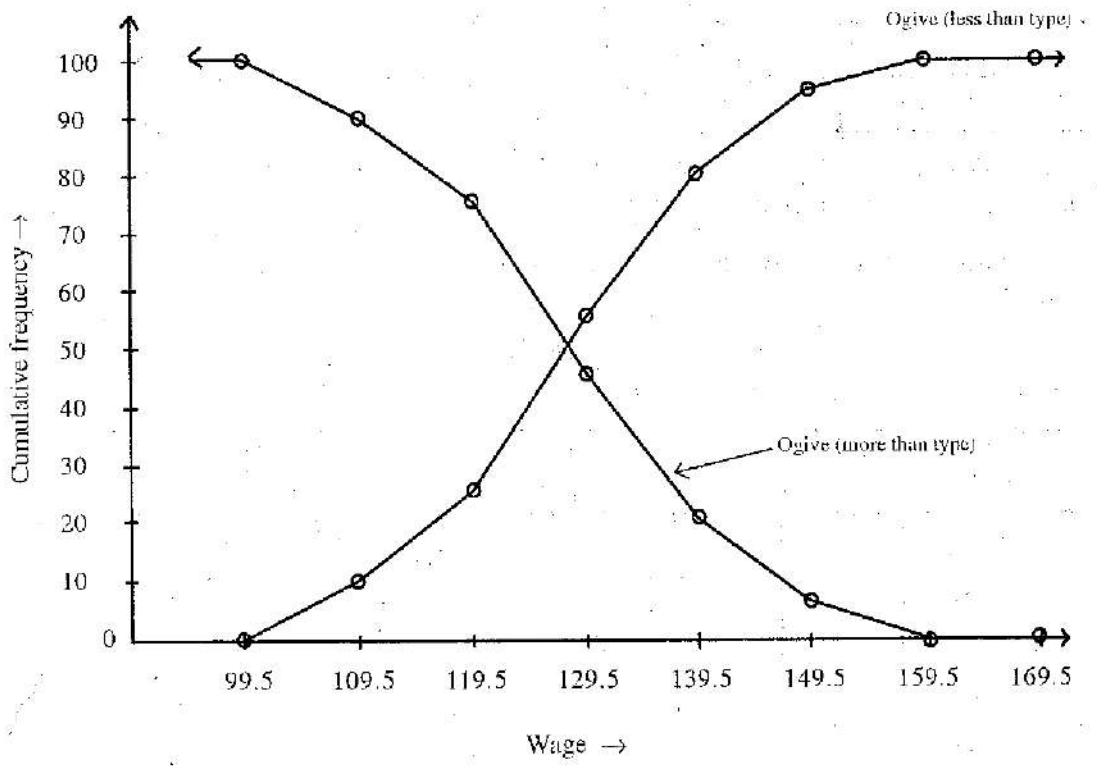


Fig. : Ogive (less than type) and ogive (more than type) of wage distribution

For ogive (less than type) the points (99·5, 0), (109·5, 10), (119·5, 25), (129·5, 55), (139·5, 80), (149·5, 95), (159·5, 100), (169·5, 100) are plotted considering wage along horizontal axis and cumulative frequency along vertical axis and the points are joined by line segments and a line parallel to horizontal axis is drawn from (169·5, 100). This ogive (less than type) is drawn. This is elongatied S type.

For ogive (more than type) the points (89·5, 100), (99·5, 100), (109·5, 90), (119·5, 75), (129·5, 45), (139·5, 20), (149·5, 5), (159·0,), (169·5, 0) are plotted and they are joined by line segments and a line, paralled to horizontal axis from (99·5, 100), is drawn to get the said curve. This is inverse clougated S type.

Example : Construct a cumulative frequency distribution (less than type) using following data :—

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	Total
No. of students	5	15	20	30	15	10	5	100

Find the number of students who secured marks between 25 and 45 from the graph and also from the table mathematically.

Solution : First construct cumulative frequency (less then type) distribution table.

Table

Cumulative frequency distribution

Class mark	No. of students (frequency)	cumulative frequency less than type
0-10	5	5
10-20	15	20
20-30	20	40
30-40	30	70
40-50	15	85
50-60	10	95
60-70	5	100

25 and 45 lies in the class 20-30 and 40-50 det x and y be cumulative frequency (less than type) corresponding to 25 and 45 consider the following table.

Table

Class mark	cum freq (less than type)
20	20
25←	→x
30	40
40	70
45←	→y
50	85

Then $\frac{25-20}{30-20} = \frac{x-20}{40-20}$ and $\frac{45-40}{50-40} = \frac{y-70}{85-70}$

i.e., $x = 20 + \frac{5}{10} \times 20 = 20 + 10 = 30$

and $y = 70 + \frac{5}{10} \times 15 = 70 + 7{\cdot}5 = 77{\cdot}5$

So frequency between 25 and 45 is $y - x = 77{\cdot}5 - 30 = 47{\cdot}5$. To get the value from the curve ogive (less than type) proceed as follows :

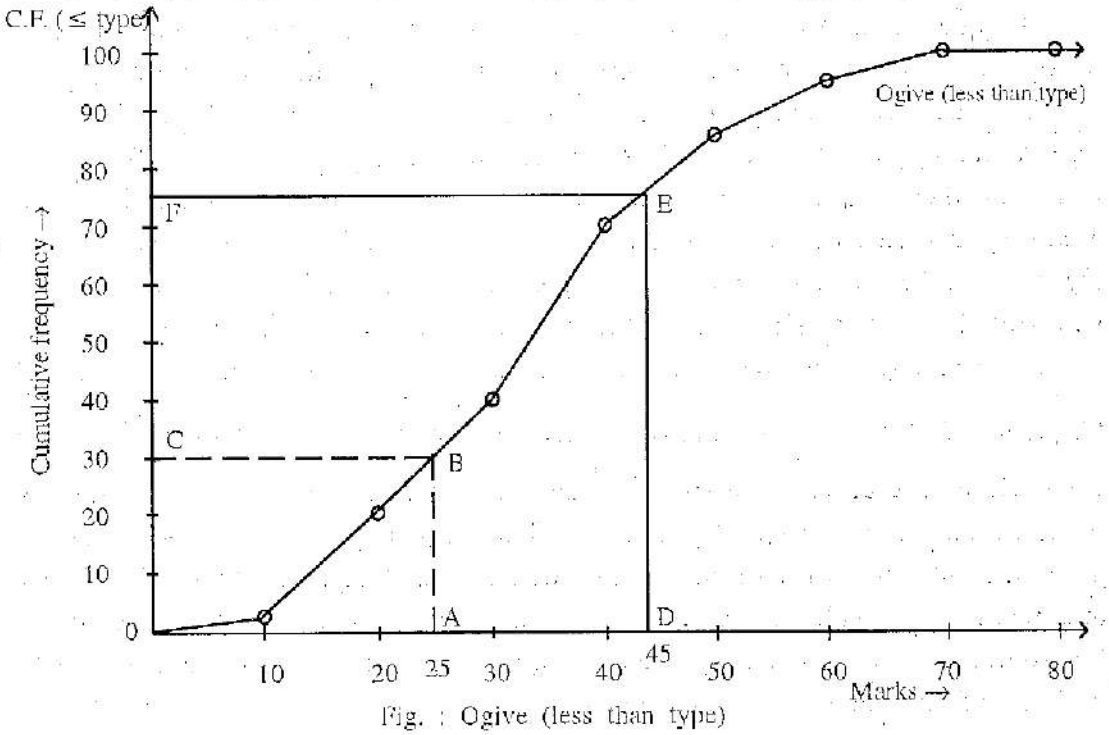


Fig. : Ogive (less than type)

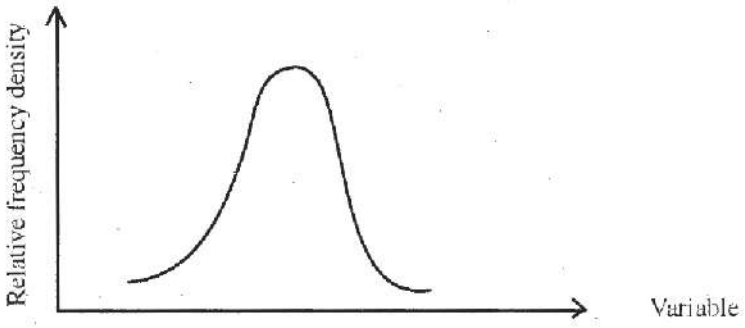
First from the cumulative frequency (less than type) distribution, consider the points (0, 0), (10, 5), (20, 20), (30, 40), (40, 70), (50, 85), (60, 95), (70, 100) and plot them in the two dimensional cartesian plan with marks along horizontal axis and cumulative frequency (less than type) along vertical axis. These points are added by line segments and get the diagram, ogive less than type which is elongated S type.

Then from A and D whose values are 25 and 45 in the horizontal axis draw perpendiculars BA and ED by dotted line and straight line to cut ogive

(less than type) at B and E respectively. Then from B and E two straight lines are drawn parallel to horizontal axis to cut vertical axis at C and F whose values are 30 and 77·5. So cumulative frequencies (less than type) for A (mark 25) and D(mark 45) are 30 and 77·5. So the frequency between 25 and 45 is 77·5 – 30 = 47·5.

2.6 Frequency curves

Frequency distribution of a continuous variable can be represented a histogram where class boundaries are taken along horizontal axis and frequency densities are taken along vertical axis and, thus the total frequency will be represented by the total area of the rectangles of histogram. If the total frequency increases (then frequency density also increases) and width of each class interval gradually decreases so that number of classes are increased, then, in the histogram, number of consecutive rectangles are increased with the decrease of width of them. In this manner if the number of rectangles are increased more and more and relative frequency densities are considered vertically instead of frequency densities, where relative frequency density = $\frac{\text{relative frequency}}{\text{class width}}$, then class widths decreases more and more and then total frequency increases rapidly though total relative frequency is unity. Thus frequency curve is conceptualized as in the limiting form of histogram, where total frequency tends to infinity and width of the classes tends to zero, and the area under the curve above horizontal axis will be unity because sum of the area of the rectangles would be sum of the relative frequences i.e., unity.

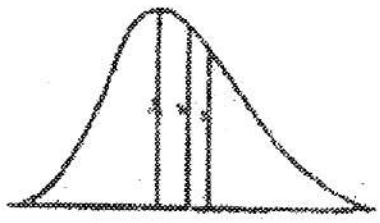


Different types of frequency curves : From the shape of histogram, the diagramatic representation of the grouped frequency distribution of a continuous variable, a good idea about the form of the distribution and the form of the frequency curve can be obtained provided the total frequency and the number classes are moderately large. The different types of distributions are broadly considered as (a) Bell shaped (i.e., unimodal) frequency distribution, (b) Multimodal distribution, (c) J-shaped distribution and (d) U-shaped distribution.

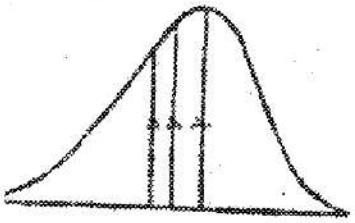
(a) **Bell shaped frequency distribution :** The frequency curve of this distribution will have a single maximum, not necessarily at the middle of the range of the variable, and is of bell shaped. Thus Bell shaped distributions again fall under sub categories :

(i) **Bell shaped symmetrical distribution :** Here frequency curve is perfectly symmetric having a maximum exactly at the middle of the range of the variable.

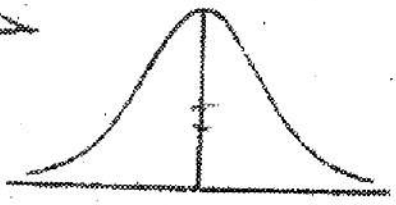
(ii) **Bell shaped moderately skew distribution :** Here frequency density dicreases at different rates on the two sides of the maximum. If the rate of decrease is faster to the left than to the right then right hand tail of the distribution will be longer than the left hand tail. Then the distribution is called positively skew. On the other hand, if the rate of decrease is faster to the right than to the left then left hand tail of the distristribution will be longer than the right hand tail. Then the distribution is called negative skew.



(a) Positive Skewness



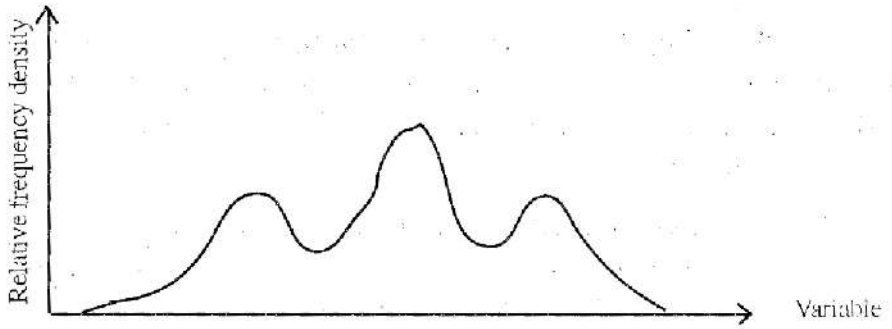
(b) Negative Skewness



(c) Zero Skewness

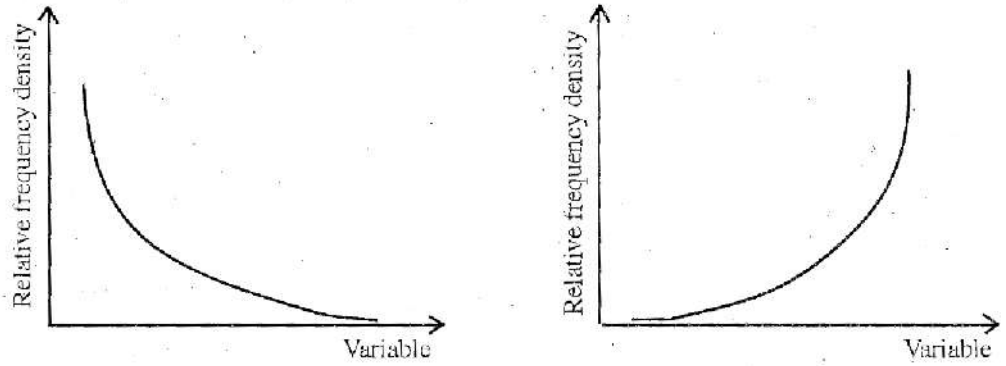
The first diagram is of frequency curve of Symmetric bellshaped distribution. Second diagram is of frequency curve of positively skewed bell shaped distribution. Third diagram is of frequency curve of negatively skewed bell shaped distribution.

(b) **Multimodal distribution :** This distribution has the frequency curve having more than one local maximum.



This is multimodal frequency curve.

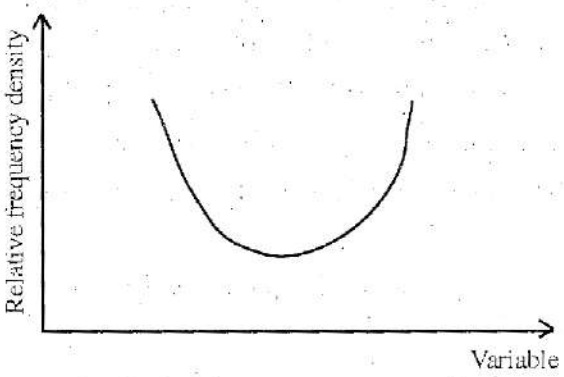
(c) **J-shaped distribution :** This extreme type of asymetric frequency distribution has its heighest frequency or frequency density at the end of the domain of the variable either at the beginning or at the end. If the first case it is called inverse J-shaped distribution and in the second case it is called J-shaped distribution.



First is the diagram of inverse J-shaped frequency curve and second is the diagram of J-shaped frequency curve.

(d) **U-shaped distribution :** For this distribution, the frequency or frequency density is lowest at the middle of the domain of the variable and

increases either at the same or different rates as the values of variable moves from middle to the left or to the right.



This is diagram of U-shaped frequency curve.

2.7 Summary

This chapter consists of frequency distribution of attributes or variables, their different components and their formations from the raw data and their graphical representations, cumulatative frequency distributions, ogives and their formations, and different types of frequency curves.

2.8 Exercises

1. What do you mean by frequency distribution of an attribute? How it can be constructed?

2. Define a variable, a discrete variable, a continuous variate with an example in each case.

3. Discuss how you would draw the frequency distribution of a discrete variable.

4. Discuss how you would draw the frequency distribution of a continuous variable.

5. The number of people in each family residing in a small town is

recorded. Discuss the procedure which one has to adopt for preparing a frequency table with these data. How are the cumulative frequency tables to be constructed?

6. What do you mean by a histogram? Explain with the help of an example in each case how it is constructed in the case of frequency distribution of a continuous variable with (i) equal class intervals, (ii) unequal class intervals.

7. Define with example (i) class limit, (ii) class boundary, (iii) class length, (iv) frequency density, (v) relative frequency, (vi) cumulative frequencies, (vii) class mark.

8. Explain how will represent a frequency distribution of a continuous variable diagramatically.

9. What are ogives? Discuss with the help of an example.

10. Write short notes on (i) column diagram, (ii) frequency polygon (iii) Histogram, (iv) cumulative frequency distribution.

11. What is a frequency curve? Write a note on different important types of frequency curves.

12. Distuiguish between an attribute and a variable with suitable illustrations.

13. Examine the characters noted below and state in each case whether the character is an attribute, a discrite variable or a continuous variable :

Annual income per family, number of letters per word, eye-colour of a person, religion of a person, height of a person in inches, life of a bulb.

14. The following is a part of the data sheet of an inspector of schools examining the student's result in a class of a school in terms of grades A, B, C, D, E in decreasing order.

B	A	B	B	A	D	C	B
B	C	D	A	A	A	B	A
A	D	B	C	B	B	A	C
D	E	E	A	D	A	C	A
C	D	D	D	C	B	A	D

Arrange the data in an appropriate frequency table. Compute the relative frequencies as well as the frequencies. Hence represent the frequency distribution in suitable diagram.

15. The following are the family sizes of 60 families in a village :

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

Draw up the frequency distribution and represent it diagramatically.

16. Draw the histogram and ogives for the following data :

Marks	:	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
Frequency	:	3	15	56	85	21	12	6	2

17. Draw the cumulative frequency diagram (less than type) for the given data. Find the median and the number of students between 42 and 52.

Marks	:	1-20	21-40	41-60	61-80	81-100
No. of students	:	9	24	43	17	7

18. Draw the histogram, ogive less than type, ogive more than type and hence derive the value of median from the following data :

Mid value of the class	:	15	25	35	45	55	65	75
Frequency	:	10	24	50	28	20	14	4

2.9 Suggested Reading

1. Gun, A. M.; Gupta, M. K. and Das gupta, B. Fundamentals of Statistics, Vol. I, World Press Pvt. Ltd., 2002.
2. Chaudhuri, S. B. Elementary Statisties, Vol. I, Shraddha Prakashani, 1986.
3. Yule, G. U. and Kendall, M. G. Introduction to the theory of statistics, charles Griffin, 1953.
4. Joydeb Sarkhel and Dutta Santosh. An Insight Into Statistics, Vol. I, Book Syndicate Pvt. Ltd., Calcutta, 1997.
5. Mills, F. C. Statistical Methods, H. Holt, 1955.

Unit 3 ❑ Central tendency

Structure

3.0 Objectives

When data are plotted one by one in a number line the values of the variable are seen to be clustered around some value, which is called central

value. Thus the characteristic of data of that they are clustering around some central value is called central tendency of data. This characteristic can be measured by a single representative value, which can be obtained by condensing the entire mass of data to a value about which all the values are distributed. This gives a comprehensive view on the entire mass of data. For example, average score of students of a school appearing in Madhyamik Examination in a certain year gives the standard of students of that school in terms of results in that year. This is obtained by totalling scores of all the students appearing in the said examination and by dividing this total by the number of appearing students in the examination. This central tendency is an important characteristic of statistical data. Its measures and their properties are considered thoroughly in this chapter.

3.1 Introduction

By the term central tendency of a given set of statistical data, we mean the characteristic of data having a central value about which observations are concentrated, since a set of data always exhibit a tendency to cluster around a specific value, called central value. For comparison of standard of living of two communities, we consider incomes per family of the communities and get two central values by averaging them. We may consider this characteristic after classification and organisation of the data into frequency distribution form. Without this characteristic we cannot compare two or more similar sets of data.

3.2 Central tendency

Generally the data has a tendency to cluster around a central value. This tendency of data is called central tendency. The central tendency is measured by a single representative value which describe the characteristic of entire distribution. This single representative value is obtained by condensing the entire mass of data to single value. This facilitates the comparison of data in two or more situations. This central value, an average of the data set, is the measure of this characteristic central tendency.

3.2.1 Criteria of a good measure

For comparing different measures of this characteristic, central tendency of data, we choose the best one among all the measures, which satisfy following criteria :

(i) It should be easily understood.

(ii) It should be simple to compute.

(iii) It should be based on all the observations.

(iv) It should not be unduly affected by extreme values.

(v) It should be rigidly defined.

(vi) It should be capable of further algebraic treatments.

(vii) It should not be affected by sampling fluctuations.

3.2.2 Different measures of central tendency

Central tendency is generally measured by three common measures :

Mean, Median and Mode.

Again means are of three types : Arithmetic mean, Geometric mean and Harmonic mean.

3.2.2.1 Arithmetic mean (or mean in short)

Most common and useful measure of central tendency is arithmetic mean. If n values of a variable x are $x_1, x_2, \ldots, x_n$ then the arithmetic mean of x is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \ldots\ldots (1)$$

For simple frequency distribution of variable x

i.e., for x_i, the i-th value of x, with frequency f_i, $i = 1, 2, \ldots, n$ and total frequency $= N = \sum_{i=1}^{n} f_i$ the arithmetic mean of x is

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{N} = \frac{1}{N}\sum_{i=1}^{n} f_ix_i \qquad \ldots\ldots (2)$$

For grouped frequency distribution of variable x

i.e., for i-th class having frequency f_i and mid value or class mark x_i,

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{N} = \frac{1}{N}\sum_{i=1}^{n} f_ix_i \qquad \ldots\ldots (3)$$

where N = total frequency of the frequency distribution.

If in a grouped frequency distribution the lower limit of the first class and the upper limit of the last class are not known it is difficult to find arithmetic mean (A.M.). When the classes (other than the first and last class) are of equal width, we may assume the widths of the open classes equal to the common width of closed classes and hence determine A.M.

From (2) and (3), (1) can be obtained when $f_1 = f_2 = \ldots = f_n = 1$ or, $f_1 = f_2 = \ldots\ldots = f_n = a$ (a constant)

Properties of Arithmetic mean (A.M.)

1. The sum of the total of the values is equal to the product of the number of values and their A.M. $(\bar{x})$ i.e., $N\bar{x} = \Sigma fx$.

2. The sum of deviations of the values from their A.M. is zero.

Proof : If $x_1, x_2, \ldots, x_n$ are the n values of the variable x or class marks of the variable x in its grouped frequency distribution with frequencies $f_1, f_2, \ldots, f_n$ respectively then $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$ are called deviations of $x_1, x_2, \ldots, x_n$ respectively from $\bar{x}$ with frequencies $f_1, f_2, \ldots, f_n$. Then algebraic sum of deviations $= \sum_{i=1}^{n} f_i(x_i - \bar{x}) = f_1(x_1 - \bar{x}) + f_2(x_2 - \bar{x}) + \ldots + f_n(x_n - \bar{x})$

$= (f_1x_1 + f_2x_2 + \ldots + f_nx_n) - N\bar{x}$ where N = total frequency $= \sum_{i=1}^{n} f_i = f_1 + f_2 + \ldots + f_n$

$= N\bar{x} - N\bar{x} = 0$

Corollary : This result can also be simplified if consider $f_1 = f_2 = \ldots = f_n = 1$

or, a (a constant) i.e., $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

3. If two variables x and y are related by the relation $y = a + bx$, then $\bar{y} = a + b\bar{x}$; where $\bar{x}$ and $\bar{y}$ are means of variables x and y and a and b are constants.

Proof : If $x_1, x_2, \ldots, x_n$ are n values of the variables x or class marks of variable x in its grouped frequency distribution with frequencies $f_1, f_2, \ldots f_n$ respectively then $y_1, y_2, \ldots, y_n$ are n values of the variable y or class marks of variable y in its corresponding grouped frequency distribution with frequencies $f_1, f_2, \ldots, f_n$ respectively where x and y are related by the relation $y = a + bx$, a, b both being constants.

Then $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{n} f_i y_i$ where N = total frequency = $\sum_{i=1}^{n} f_i$.

$$\text{So } \bar{y} = \frac{1}{N}\sum_{i=1}^{n} f_i y_i = \frac{1}{N}\sum_{i=1}^{n} f_i(a + bx_i) = \frac{1}{N}\left[\sum_{i=1}^{n} f_i a + \sum_{i=1}^{n} b f_i x_i\right]$$

$$= \frac{a}{N}\sum_{i=1}^{n} f_i + \frac{b}{N}\sum_{i=1}^{n} f_i x_i = \frac{aN}{N} + b\bar{x}$$

$$= a + b\bar{x}$$

Corollary : This proof can also be simplified if we consider $f_1 = f_2 = \ldots = f_n = 1$ or a (constant).

4. Arithmetic mean depends on the change of origin and also on change of scale.

Proof : Let $u_i = \frac{x_i - c}{d}$ where c and d are constants and origin of variable x is considered at c and d unit of x is equal to 1 unit of u, $i = 1, 2, \ldots, n$ are the n values of variable u like $u_1, u_2, \ldots, u_n$ obtained in terms of corresponding n values of variable x like $x_1, x_2, \ldots, x_n$ in order. Let the corresponding frequencies be $f_1, f_2, \ldots, f_n$ respective for the simple or grouped frequency distribution.

Then $x_i = c + du_i$ occur with frequency f_i for $i = 1, 2, \ldots, n$ and $N =$ total frequency $= \sum_{i=1}^{n} f_i$.

So, $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \frac{1}{N}\sum_{i=1}^{n} f_i(c + du_i) = \frac{cN}{N} + \frac{d\sum f_i}{N}$

$= c + d\bar{u}$ i.e., $\bar{u} = \frac{\bar{x} - c}{d}$

So mean depends on the change of origin and change of scale.

Corollary : This proof can also be simplified if we consider $f_1 = f_2 = \ldots = f_n = 1$ or a(constant).

5. If a group of N_1 values has arithmetic mean $\bar{x}_1$ and another group of N_2 values has arithmetic mean $\bar{x}_2$, then arithmetic mean of the composite group (i.e., the two groups combined) of $N_1 + N_2$ values is

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} \qquad \ldots\ldots (4)$$

Proof : Let the first group contains values $x_{11}, x_{12}, \ldots, x_{1n_1}$ with frequencies $f_{11}, f_{12}, \ldots, f_{1n_1}$ respectively and the sceond group contains values $x_{21}, x_{22}, \ldots, x_{2n_2}$ with frequencies $f_{21}, f_{22}, \ldots, f_{2n_2}$ respectively and $N_1 = \sum_{i=1}^{n_1} f_{1i}$, $N_2 = \sum_{j=1}^{n_2} f_{2j}$.

When the two groups are combined together the combined mean $\bar{x}$ is calculated as

$$\bar{x} = \frac{f_{11}x_{11} + f_{12}x_{12} + \ldots + f_{1n_1}x_{1n_1} + f_{21}x_{21} + f_{22}x_{22} + \ldots + f_{2n_2}x_{2n_2}}{N_1 + N_2}$$

$$= \frac{(f_{11}x_{11} + f_{12}x_{12} + \ldots + f_{1n_1}x_{1n_1}) + (f_{21}x_{21} + f_{22}x_{22} + \ldots + f_{2n_2}x_{2n_2})}{N_1 + N_2}$$

$$= \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2}$$

since $\bar{x}_1 = \frac{1}{N_1} = \sum_{i=1}^{n_1} f_{1i}x_{1i}$ and $\bar{x}_2 = \frac{1}{N_2} = \sum_{j=1}^{n_2} f_{2j}\ x_{2j}$.

Corollary 1 : This proof can also be simplified if we consider $f_{11} = f_{12} = \ldots = f_{1n_1} = 1$ or a (constant) and $f_{21} = f_{22} = \ldots = f_{2n_2} = 1$ or a(constant).

Then $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ where $\bar{x_1} = \dfrac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}$ and $\bar{x_2} = \dfrac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}$

Corollary 2. If there are k groups of $N_1, N_2, \ldots, N_k$ values with means $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_k$ respectively then mean of the composite group is

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + \ldots + N_k\bar{x}_k}{N_1 + N_2 + \ldots + N_k}$$

Corollary 3 : For two group means $\bar{x}_1$ and $\bar{x}_2$ with N_1 and N_2 observations the grouped mean $\bar{x}$ lies between $\bar{x}_1$ and $\bar{x}_2$.

Proof : Let $\bar{x}_1 \leq \bar{x}_2$.

Then $\bar{x} - \bar{x}_1 = \dfrac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} - \bar{x}_1 = \dfrac{N_1\bar{x}_1 + N_2\bar{x}_2 - N_1\bar{x}_1 + N_2\bar{x}_1}{N_1 + N_2}$

$$= \frac{N_2\bar{x}_2 - N_2\bar{x}_1}{N_1 + N_2} = \frac{N_2(\bar{x}_2 - \bar{x}_1)}{N_1 + N_2} \geq 0.$$

or, $\bar{x} \geq \bar{x}_1$ since $\bar{x}_2 \geq \bar{x}_2$.

Again $\bar{x}_2 - \bar{x} = \bar{x}_2 - \dfrac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} = \dfrac{N_1\bar{x}_2 + N_2\bar{x}_2 - N_1\bar{x}_1 - N_2\bar{x}_2}{N_1 + N_2}$

$$= \frac{N_1\bar{x}_2 - N_1\bar{x}_1}{N_1 + N_2} = \frac{N_1(\bar{x}_2 - \bar{x}_1)}{N_1 + N_2} \geq 0 \text{ since } \bar{x}_2 \geq \bar{x}_1$$

or, $\bar{x}_2 \geq \bar{x}$.

Thus if $\bar{x}_1 \leq \bar{x}_2$, $\bar{x}_1 \leq \bar{x} \leq \bar{x}_2$.

Again similarly it can be shown that if $\bar{x}_2 \leq \bar{x}_1$ then $\bar{x}_2 \leq \bar{x} \leq \bar{x}_1$.

Thus from these result we can pvove that lies $\bar{x}$ lies between $\bar{x}_1$ and $\bar{x}_2$.

Merits : (i) It is easy to calculate and simple to understand.

(ii) It is based on all the observations.

(iii) It is capable of further mathematical treatments.

(iv) It is rigidly defined and is least affected by sampling fluctations.

(v) It can easily be used to compare two or more frequency distributions.

(vi) It does not necessitate arrangement of data.

Demerits : (i) It is affected by very large or very small values.

(ii) It cannot be determined in case of grouped frequency distribution with any open ended class.

3.2.2.2 Geometric mean (*G*).

The geometric mean (G) of n positive values $x_1, x_2, \ldots\ldots x_n$ is the n-th root of the product of the values i.e.,

$$G = \sqrt[n]{x_1 x_2 \ldots x_n} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

G can easily be calculated from the relation,

$$\log G = \frac{1}{n} \log \left(\prod_{i=1}^{n} x_i \right) = \frac{1}{n} \sum_{i=1}^{n} \log x_i.$$

since $\log(x_1 \; x_2 \ldots x_n) = \log x_1 + \log x_2 + \ldots + \log x_n$.

and $G = \text{anti-log} \left[\frac{1}{n} \sum_{i=1}^{n} \log x_i \right]$

For simple frequency distribution or grouped frequency distrubution, where x_i is the i-th value or class mark of i-th class with frequiency f_i for $i = 1, 2, \ldots, n$, the geometric mean G is

$$G = \left(x_1^{f_1} x_2^{f_2} \ldots x_n^{f_n} \right)^{\frac{1}{N}} \text{ where } N = \text{total frequency} = \sum_{i=1}^{n} f_i.$$

$$= \left(\prod_{i=1}^{n} x_i^{f_i} \right)^{\frac{1}{N}}$$

This geometric mean can be calculated from $\log G = \frac{1}{N} \sum_{i=1}^{n} f_i \log x_i$.

Important Properties :—

1. The product of all the values is equal to the N-th power of their *G.M.* (G).

i.e., $x_1^{f_1} x_2^{f_2} \ldots x_n^{f_n} = G^N$ in case of simple of grouped frequency distribution and $x_1 x_2 \ldots x_n = G^n$ when all frequencies are equal.

2. The logarithm of *G.M.* of N observations is equal to the A.M. of logarithms of N observations.

i.e., $\log N = \frac{1}{N} \sum_{i=1}^{n} f_i \log x_i$ in case of simple or grouped frequcney distributions

and $\log G = \frac{1}{n} \sum_{i=1}^{n} \log x_i$ when all frequencies are equal.

3. If two variables x and y are related by the relation $y = bx$ then geometric mean of x (G_x) and geometric mean of y (G_y) are related by the relation $G_y = bG_x$ where b is positive constant.

Proof : Let variable x takes value x_i with frequency f_i for $i = 1, 2, \ldots, n$ and $N = \sum_{i=1}^{n} f_i$. Then the variable y also takes the value y_i with frequency f_i for $i = 1, 2, \ldots, n$. Then $y_i = bx_i$, $i = 1, 2, \ldots, n$, and

$$G_y = \left(\prod_{i=1}^{n} y_i^{f_i} \right)^{\frac{1}{N}} = \left(\prod_{i=1}^{n} (bx_i)^{f_i} \right)^{\frac{1}{N}} = \left(\prod_{i=1}^{n} b^{f_i} xi^{f_i} \right)^{\frac{1}{N}}$$

$$= \left(b^{\sum_{i=1}^{n} f_i} \prod_{i=1}^{n} x_i^{f_i} \right)^{\frac{1}{N}} = \left(b^N \prod_{i=1}^{n} x_i^{f_i} \right)^{\frac{1}{N}} = b^{\frac{N}{N}} \left(\prod_{i=1}^{n} xi^{f_i} \right)^{\frac{1}{N}} = bG_x$$

4. For two groups with number of observations N_1 and N_2 and geometric means G_1 and G_2 respectively, the geometric mean G of the combined group is

$$G = \left(G_1^{N_1} \; G_2^{N_2} \right)^{\frac{1}{N_1 + N_2}}$$

Proof : Suppose for two frequency distributions with total frequencies N_1 and N_2, the first set of observations are $x_{11}, x_{12}, \ldots, x_{1n_1}$, with frequences $f_{11}, f_{12}, \ldots, f_{1n_1}$ respectively and the second set of observations are $x_{21}, x_{22}, \ldots, x_{2n_2}$

with frequencies $f_{21}, f_{22}, \ldots, f_{2n_2}$ respectively so that $N_1 = \sum_{i=1}^{n_1} f_{1i}$, $N_2 = \sum_{j=1}^{n_2} f_{2j}$.

The combined geometric mean G can be obtained in terms of group geometric means G_1 and G_2 of first and second group in following way.

$$G = \left(x_{11}^{f_{11}} x_{12}^{f_{12}} \cdots x_{1n_1}^{f_{1n_1}} x_{21}^{f_{21}} x_{22}^{f_{22}} \cdots x_{2n_2}^{f_{2n_2}} \right)^{\frac{1}{N_1+N_2}}$$

$$= \left(\prod_{i=1}^{n_1} x_{1i}^{f_{1i}} \prod_{j=1}^{n_2} x_{2j}^{f_{2j}} \right)^{\frac{1}{N_1+N_2}}$$

$$= \left(G_1^{N_1} G_2^{N_2} \right)^{\frac{1}{N_1+N_2}} \text{ since } G_1 = \left(\prod_{i=1}^{n_1} x_{1i}^{f_{1i}} \right)^{\frac{1}{N_1}} \text{ and}$$

$$G_2 = \left(\prod_{j=1}^{n_2} x_{2j}^{f_{2j}} \right)^{\frac{1}{N_2}}$$

Generalisation : For k groups with geometric means $G_1, G_2, \ldots, G_k$ and corresponding total frequencies $N_1, N_2, \ldots, N_k$ the grouped geometric mean (G) is

$$G = \left(G_1^{N_1} G_2^{N_2} \ldots G_k^{N_k} \right)^{\frac{1}{N_1+N_2+\ldots+N_k}}$$

Corollary : If all the distinct values occur once in each of two groups then first group has n_1 values $x_{11}, x_{12}, \ldots x_{1n_1}$ with G.M. G_1 and second group has n_2 values $x_{21}, x_{22}, \ldots x_{2n_2}$, with G.M. G_2 and then combined geometric mean G can be obtained from the relation $G = \left(G_1^{n_1} G_2^{n_2} \right)^{\frac{1}{n_1+n_2}}$.

Merits : (i) It is based on all observations and capable of further algebraic treatments.

(ii) It is rigidly defined and it gives less weight to larger items and more weight to smaller items than the arithmetic mean.

(iii) It is useful in average ratios and percentages in determining ratio of change.

(iv) When the observations are in geometric progression geometric mean is the suitable average.

Demerits : (i) It is difficult to compute and to interpret. So it is of less use.

(ii) If any value of a set of values be zero or negative, then geometric mean cannot be determined.

3.2.2.3 Harmonic mean (H)

The harmonic mean (H) for n observations $x_1, x_2, \ldots, x_n$ is the number of observations divided by the sum of the reciprocals of observations i.e.,

$$H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Also $\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$

i.e., inverse of harmonic mean is the arithmetic mean of the reciprocals of observations.

For a frequency distribution, where variable x takes values $x_1, x_2, \ldots, x_n$ with frequencies $f_1, f_2, \ldots, f_n$ respectively, the harmonic mean H is

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

Important properties : 1. Inverse of harmonic mean is the arithmetic mean of inverses of values.

i.e. $\frac{1}{H} = \frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{x_i}$

2. For two variables x and y related by $y = bx$, H_y = harmonic mean of variable y, H_x = harmonic mean of variable x,

$$H_y = bH_x$$

Proof : Let the variable x takes values $x_1, x_2, \ldots, x_n$ with frequencies $f_1, f_2, \ldots, f_n$ respectively. Then the variable y takes the values $y_1 = bx_1, y_2 = bx_2, \ldots, y_n = bx_n$ with frequencies $f_1, f_2, \ldots, f_n$ respectively and N = total frequency $= \sum_{i=1}^{n} f_i$. Then

$$H_y = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{y_i}}, \quad H_x = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

$$\text{Thus } H_y = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{bx_i}} = \frac{bN}{\sum_{i=1}^{n} \frac{f_i}{x_i}} = bH_x$$

2. For two groups with number of observation N_1 and N_2 with harmonic means H_1 and H_2 respectively, the grouped harmonic mean H is obtained from the relation

$$\frac{N_1 + N_2}{H} = \frac{N_1}{H_1} + \frac{N_2}{H_2}.$$

Proof : Let one frequency distributions has values $x_{11}, x_{12}, \ldots, x_{1n_1}$ with frequencies $f_{11}, f_{12}, \ldots, f_{1n_1}$ respectively with harmonic mean H_1 and other frequency frequency distribution has values $x_{21}, x_{22}, \ldots, x_{2n_2}$ with frequencies $f_{21}, f_{22}, \ldots, f_{2n_2}$ rewpectively with harmonic mean H_2.

$$\text{Let } N_1 = \sum_{i=1}^{n_1} f_{1i}, \quad N_2 = \sum_{j=1}^{n_2} f_{2j}$$

The grouped harmonic mean (H) and harmonic mean of i-th group (H_i) for $i = 1, 2$ are

$$H = \frac{N_1 + N_2}{\sum_{i=1}^{n_1} \frac{f_{1i}}{x_{1i}} + \sum_{j=1}^{n_2} \frac{f_{2j}}{x_{2j}}} \quad H_1 = \frac{N_1}{\sum_{i=1}^{n_1} \frac{f_{1i}}{x_{1i}}} \text{ and } H_2 = \frac{N_2}{\sum_{i=1}^{n_2} \frac{f_{2j}}{x_{2j}}}$$

$$\text{Then } \frac{N_1 + N_2}{H} = \sum_{i=1}^{n_1} \frac{f_{1i}}{x_{1i}} + \sum_{i=1}^{n_2} \frac{f_{2j}}{x_{2j}} = \frac{N_1}{H_1} + \frac{N_2}{H_2}$$

Generalisation : For k groups with harmonic means $H_1, H_2, \ldots, H_k$ and

total frequences N_1, N_2, ..., N_k respectively, the grouped harmonic mean H can be obtained from

$$\frac{N_1 + N_2 + \ldots + N_k}{H} = \frac{N_1}{H_1} + \frac{N_2}{H_2} + \ldots + \frac{N_k}{H_k}$$

Merits : (i) Like A.M. and G.M., it is also based on all observations.

(ii) It is capable of algebraic treatments.

(iii) In problems of getting average speed relating to time, distance and speed it gives appropriate result.

Demerits : (i) It is neither easily understood nor easy to calculate.

(ii) If any value of a set of observation is zero, harmonic mean cannot be determined.

(iii) It gives largest weight to smallest value and smallest weight to largest value. This is not desirable generally and as such it is not of much use in analysis of economic data.

3.2.2.4 Relationship between A.M., G.M. and H.M.

Theorem : For n positive real observations

A.M. $\geq$ G.M. $\geq$ H.M.

Proof : For two positive and real observations x_1 and x_2,

$$\left(\sqrt{x_1} - \sqrt{x_2}\right)^2 \geq 0 \quad \text{or} \quad x_1 + x_2 - 2\sqrt{x_1 x_2} \geq 0$$

or, $$\frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2} \qquad \ldots\ldots (1)$$

So, for $n = 2$, A.M. $\geq$ G.M.

Thus for four positive and real observations x_1, x_2, x_3 and x_4

$$\left(\sqrt{\frac{x_1 + x_2}{2}} - \sqrt{\frac{x_3 + x_4}{2}}\right)^2 \geq 0$$

or, $$\frac{x_1 + x_2}{2} + \frac{x_3 + x_4}{2} \geq 2\sqrt{\frac{x_1 + x_2}{2}}\sqrt{\frac{x_3 + x_4}{2}}$$

by (1) $\frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$ and $\frac{x_3 + x_4}{2} \geq \sqrt{x_3 x_4}$

Thus $\frac{x_1 + x_2 + x_3 + x_4}{2} \geq 2\sqrt{\sqrt{x_1 x_2}\sqrt{x_3 x_4}}$

or, $\frac{x_1 + x_2 + x_3 + x_4}{4} \geq \sqrt[4]{x_1 x_2 x_3 x_4}$

So for $n = 4$, A.M. $\geq$ G.M.

Proceeding this way we can show A.M. $\geq$ G.M. for $n = 2, 4, 8, 16$ etc. So when $n = 2^m$, m being a positive integer, A.M. $\geq$ G.M. for n positive real observations $x_1, x_2, \ldots, x_n$.

Consider $2^{m-1} < n < 2^m$ i.e., n lies between 2^{m-1} and 2^m but $n \neq 2^{m-1}$ or 2^m. Then consider $A = \frac{x_1 + x_2 + \ldots + x_n}{n}$ = A.M. of n given observations. and $N = 2^m$.

Consider N real values of which first n are $x_1, x_2, \ldots, x_n$ and the last $(N - n)$ are all equal to A.

Then for $N = 2^m$ observations A.M. $\geq$ G.M.

Then $\frac{x_1 + x_2 + \ldots + x_n + A + A + \ldots + A}{n + (N - n)} \geq (x_1 x_2 \ldots x_n \text{ A.A. } \ldots A)^{1/N}$

or, $\frac{nA + (N - n)A}{N} \geq (x_1 x_2 \ldots x_n . A^{N-n})^{1/N}$

or, $A \geq (G^n A^{N-n})^{1/N}$

where G = geometric mean of $x_1, x_2, \ldots, x_n = (x_1 x_2 \ldots x_n)^{1/n}$

i.e., $G^n = x_1 x_2 \ldots x_n$.

Thus $A^N \geq G^n . A^{N-n}$

or, $\frac{A^N}{A^{N-n}} \geq G^n$

or, $A^{N-N+n} \geq G^n$

or, $A^n \geq G^n$

or, $A \geq G$.

So when $2^{m-1} < n < 2^m$, A.M. $\geq$ G.M. Hence A.M. $\geq$ G.M. for any number of positive and real observations. ... (2)

As $x_1, x_2, \ldots, x_n$ are real and positive, $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}$ are also real and positive. Then for any positive integral values of n, A.M. $\geq$ G.M. for the observations $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}$.

Thus $$\frac{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}}{n} \geq \left(\frac{1}{x_1}\cdot\frac{1}{x_2}\cdots\frac{1}{x_n}\right)^{\frac{1}{n}} = \frac{1}{(x_1x_2\ldots x_n)^{\frac{1}{n}}}$$

Then reversing, $$\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}} \leq (x_1x_2\ldots x_n)^{\frac{1}{n}}$$

or H.M. $\leq$ G.M.

Thus G.M. $\geq$ H.M. for n real and positive observation. ..... (3)

combining these two results (2) and (3),

A.M. $\geq$ G.M. $\geq$ H.M.

holds for n neal and positive observations.

3.2.2.5 Uses of A.M., G.M. and H.M.

(1) Distance covered = speed × time taken

Then time = $\frac{\text{distance}}{\text{speed}}$

If k distances $d_1, d_2, \ldots, d_k$ are covered with k different speeds $s_1, s_2, \ldots, s_k$ respectively then

$$\text{average speed} = \frac{\text{total distance covered}}{\text{total time taken}} = \frac{d_1+d_2+\ldots+d_k}{\frac{d_1}{s_1}+\frac{d_2}{s_2}+\ldots+\frac{d_k}{s_k}}$$

i.e., average speed is the weighted harmonic mean of k speeds with distances as weights. Here time is unknown. If distance is unknown i.e., k distances are covered with speeds $s_1, s_2, \ldots, s_k$ in times $t_1, t_2, \ldots, t_k$ respectively then

$$\text{average speed} = \frac{\text{total distance covered}}{\text{total time taken}} = \frac{t_1s_1+t_2s_2+\ldots+t_ks_k}{t_1+t_2+\ldots+t_k}$$

i.e., average speed is the weighted arithmetic mean of k speeds with times as weights.

It speeds are unknown i.e., k distances $d_1, d_2, \ldots, d_k$ are covered in times $t_1, t_2, \ldots, t_k$ respectively, then

$$\text{average speed} = \frac{\text{total distance covered}}{\text{total time taken}} = \frac{d_1 + d_2 + \ldots + d_k}{t_1 + t_2 + \ldots + t_k}$$

(2) If k types of milk of amount $m_1, m_2, \ldots, m_k$ litres are sold at rates $r_1, r_2, \ldots, r_k$ Rs per litre.

Then average rate of price in which it is sold is

$$\frac{\text{total selling price}}{\text{total quantity sold}} = \frac{m_1 r_1 + m_2 r_2 + \ldots + m_k r_k}{m_1 + m_2 + \ldots + m_k}$$

It is weighted arithmetic mean of rates of selling price weights being quantity of milk sold. Here amount of money in which the differents milks are sold are unknown. If amount of milks are unknown i.e., k types of milk are sold at prices $P_1, P_2, \ldots, P_k$ Rs at the rates $r_1, r_2, \ldots r_k$ rupees per litre then average selling price is,

$$\frac{\text{Total seling price}}{\text{Total quantity sold}} = \frac{P_1 + P_2 + \ldots + P_k}{\frac{P_1}{r_1} + \frac{P_2}{r_2} + \ldots + \frac{P_k}{r_k}}$$

This is weighted harmonic mean of rates of selling price with weights as the money at which they are sold.

(3) Geometric mean is most frequently used in the determination of average percentage of change. For example (a) if the population increases at the rates $r_1, r_2, \ldots, r_k$ percents per year in first n_1 years, next n_2 years, and in last n_k years then average rate of increase $r\%$ is obtained from

$$P_0\left(1+\frac{r}{100}\right)^{n_1+n_2+\ldots+n_k} = P_0\left(1+\frac{r_1}{100}\right)^{n_1}\left(1+\frac{r_2}{100}\right)^{n_2}\ldots\left(1+\frac{r_k}{100}\right)^{n_k}$$

where P_0 = initial population.

or, $$\text{from}\left(1+\frac{r}{100}\right)^{n_1+n_2+\ldots+n_k} = \left(1+\frac{r_1}{100}\right)^{n_1}\left(1+\frac{r_2}{100}\right)^{n_2}\ldots\left(1+\frac{r_k}{100}\right)^{n_k}$$

or, $$\text{from } (n_1 + n_2 + \ldots + n_k)\log\left(1+\frac{r}{100}\right) = \sum_{i=1}^{k} n_i \log\left(1+\frac{r_i}{100}\right)$$

(b) If a machine depreciates in k consecutive periods of length n_1, n_2, ..., n_k years at the rates r_1, r_2, ..., r_k percents per, year then average rate of depreciation $r\%$ can be obtained from

$$\left(1-\frac{r}{100}\right)^{n_1+n_2+\ldots+n_k} = \left(1-\frac{r_1}{100}\right)^{n_1}\left(1-\frac{r_2}{100}\right)^{n_2} \ldots 1-\left(\frac{r_k}{100}\right)^{n_k}$$

or, $(n_1 + n_2 + \ldots + n_k) \log\left(1-\frac{r}{100}\right) = \sum_{i=1}^{k} n_i \log\left(1-\frac{r_i}{100}\right)$.

3.2.2.5 Median

Median of a set of values is the middle most value when the values are arranged either in non-decreasing or in non-increasing order. If n, the number of observations is odd then median is the middle most observation i.e., $\frac{n+1}{2}$ -th value in the ordered data. If n is even then median is the average of two middlemost observations i.e., average of $\frac{n}{2}$-th and $\left(\frac{n}{2}+1\right)$-th value of ordered data.

For the simple frequency distribution of a variable x first determine the cumulative frequencies (C.F.'s) F of less than type as follows :

Value of variable (x)	x_1	x_2	x_3	...	x_n	Total
frequency (f)	f_1	f_2	f_3	...	f_n	N
C.F. (less than type) (F)	F_1	F_2	F_3	...	F_n	—

Here $F_1 = f_1$, $F_2 = f_1 + f_2$, $F_3 = f_1 + f_2 + f_3$, ..., $N = F_n = f_1 + f_2 + \ldots + f_n$ = total frequency. If N is odd, then median is that x for which least $F \geq \frac{N+1}{2}$. If N is even, then median is the average of two x's for which least $F \geq \frac{N}{2}$ and least $F \geq \frac{N}{2} + 1$ respectively.

For a grouped frequency distribution of a variable x, first determine the cumulative frequencies (C.F.) F (less than type) as follows :

Class of variable (x)	$L_1 - U_1$	$L_2 - U_2$	...	$L_n - U_n$	Total
frequency (f)	f_1	f_2	...	f_n	N
C.F. (less than type) (F)	F_1	F_2	...	$F_n = N$	

where $F_1 = f_1$, $F_2 = f_1 + f_2$, $F_3 = f_1 + f_2 + f_3$, ...

$F_n = f_1 + f_2 + \ldots + f_n = N$ = Total frequency and $L_i - U_i$ is the i-th class in which L_i = lower boundary and U_i = upper boundary of the i-th class and $U_1 = L_2$, $U_2 = L_3$, ..., $U_{n-1} = L_n$.

Median class is the class in which median lies i.e., for that class least $F \geq \frac{N}{2}$. Let $L - U$ be the median class. Then

$$M_e = \text{median} = L + \frac{\frac{N}{2} - F}{f_m} \times c$$

where L = lower boundary of the median class

F = cumulative frequency of the preceding class to the median class

f_m = frequency of median class

c = class length of median class (i.e., $c = U - L$).

Important property. If y is a linear function of x i.e., $y = c + dx$, then the median of the two variable are related as $M_e(y) = c + dM_e(x)$ where $M_e(x)$ and $M_e(y)$ are the medians of variables x and y, c and d being constants.

Proof : Let n values $x_1, x_2, \ldots, x_n$ of variable x be such that $x_1 \leq x_2 \leq \ldots \leq x_n$

Thus $y_i = c + dx_i$ obtained by putting $y = y_i$ and

$x = x_i$ in the relation $y = c + dx_i$

Then $y_1 \leq y_2 \leq \ldots \leq y_n$ when $d > 0$

and $y_1 \geq y_1 \geq \ldots \geq y_n$ when $d < 0$

If x_k is the middle most value of the variable x, so is y_k for the variable y. So

$$M_e(y) = c + dM_e(x).$$

Similarly if there are two middlemost values of variable x, say x_k and x_{k+1} then the corresponding two middlemost values of variable y are $y_k = c + dx_k$ and $y_{k+1} = c + dx_{k+1}$.

So the median of y and that of x will be $M_e(y) = \frac{1}{2}(y_k + y_{k+1})$, $M_e(x) = \frac{1}{2}(x_k + x_{k+1})$.

Thus $M_e(y) = \frac{1}{2}(y_k + y_{k+1}) = \frac{1}{2}[(c + dx_k) + (c + dx_{k+1})]$

$$= \frac{c+c}{2} + d\left(\frac{x_k + x_{k+1}}{2}\right)$$

$$= c + dM_e(x).$$

Generalised result : If $y = h(x)$ be a monotonic function of x, then $M_e(y) = h(M_e(x))$.

Merits : (i) It is rigidly defined and easily understandable.

(ii) It is simple to compute.

(iii) It is not affected by the presence of few extremely large or extremely small values.

(iv) It can be calculated even if one or both the terminal classes of a grouped frequency distribution are open,

(v) It can be determined graphically by drawing ogives or can be located by inspection after arranging the data in order either increasing or decreasing.

Demerits : (i) Median is not based on all observations.

(ii) It is not amenable to algebraic treatments

(iii) It may not be uniquely determined in case of even number of observations.

(iv) Calculation of median in case of continuous variable is difficult

(v) Median is affected more by sampling fluctuations than the mean.

3.2.2.7 Mode

Mode is the value of the variable which occurs most frequently. It is the

value of the variable which has highest frequency. There may exist more than one mode in a frequency distribution. Then the distribution is multimode.

In the frequency distribution of a discrete variable the mode can be obtained by observing highest frequency.

In case of a continuous variable the class having highest frequency density is called modal class. Let it be $L - U$ where L = lower boundary and U = upper boundary of the modal class. Let C_{m_0} = class length of modal class = $U - L$ and f_{m_0} be the frequency of modal class.

Let $C_{m_0 - 1}$ and $f_{m_0 - 1}$ be the class length and frequency of the proceeding class to the modal class.

Let $C_{m_0 + 1}$ and $f_{m_0 + 1}$ be the class length and frequency of the proceeding class to the modal class.

Then

$$\text{Mode } M_0 = L + \frac{\dfrac{f_{m_0}}{c_{m_0}} - \dfrac{f_{m_0-1}}{c_{m_0-1}}}{\dfrac{f_{m_0}}{c_{m_0}} - \dfrac{f_{m_0-1}}{c_{m_0-1}} - \dfrac{f_{m_0+1}}{c_{m_0+1}}} \times c_{m_0} \quad Cm_0$$

If three consecutive classes including modal class in the middle are of equal length then for modal class $L - U$ of length C where L = lower boundary and U = upper boundary, Mode M_0 can be determined from the relation

$$M_0 = L + \frac{f_{m_0} - f_{m_0-1}}{2f_{m_0} - f_{m_0-1} - f_{m_0+1}} \times c$$

where f_{m_0}, $f_{m_0 - 1}$ and $f_{m_0 + 1}$ are the frequencies of modal class, the class just preceeding the modal class and the class just proceeding to the modal class.

Important properties : If two variables x and y are linearly related and connected by the relation $y = a + bx$, a and b being constants then mode of $x = M_0$ and mode of y will be $a + bM_0$.

Proof : In case of simple frequency distribution if M_0 be the mode of variable x, then M_0 has highest frequency. Then ultimately $a + bM_0$ also has highest frequency for the variable y. Thus then mode of $y = a + b$ (Mode of x).

In case of grouped distribution, let $L_0 - U_0$ be the modal class of variable x then $L'_0 - U'_0$ will be the modal class of variable y, where $L'_0 = a + bL_0$ and $U'_0 = a + bU_0$. Thus $U'_0 - L'_0 = b(U_0 - L_0) = bc_0$ where c_0 is the class length of modal class of x.

Let f_0 be the frequency of modal class of variable x. We consider a grouped frequency distribution which has three consecutive classes of equal length c_0 (say) with modal class at the centre for the variable x, and f_{-1} and f_1 are the frequencies of the classes just preceding and proceeding to modal classes.

$$\text{Then mode of } y = L'_0 + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times bc_0$$

$$= a + bL_0 + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times bc_0$$

$$= a + b\left(L_0 + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times c_0\right)$$

$$= a + b \times \text{Mode of } x$$

Thus $M_0(y) = a + bM_0(x)$

Where $M_0(y)$ and $M_0(x)$ are the modes of the variables y and x respectively. The result can similarly be proved for unequal class intervals for the variable x.

Merits : (i) The mode is rigidly defined and simple to compute

(ii) Significance of mode is easily comprehensible.

(iii) It is not affected by the presence of a few extreme values.

(iv) It can be calculated even for a grouped frequency distribution having one or both of the terminal classes open.

Demerits : (i) It is not based on all the observations.

(ii) It is not amenable to algebraic treatments.

(iii) Determination of exact value of mode may not be possible.

3.2.2.8 Other Positional Measures

Besides median there are measures, which divide an ordered set of data into equal parts. Among these, the most important measures are quartites, deciles

and percentiles. Quartiles are those values which divide the nondecreasing ordered data into 4 equal parts by three quartiles Q_1, Q_2 and Q_3 the first, second and third quartile, corresponding to which the cumulative frequencies (less than type) are $\frac{N}{4}, \frac{2N}{4}$ and $\frac{3N}{4}$ respectively, N being total frequency when the values of the variables are arranged in nondecreasing order. There are 9 deciles D_1, D_2, ..., D_9 i.e., first, second, ninth decile for whom cumulative frequencies (less than type) are $\frac{N}{10}, \frac{2N}{10}, \ldots, \frac{9N}{10}$ respectively and divide the nondecreasing ordered set of data into 10 equal parts. There are 99 percentiles P_1, P_2, ..., P_{99} i.e., first, second, ..., 99-th percentiles for whom the cumulative frequencies (less than type) are $\frac{N}{100}, \frac{2N}{100}, \ldots, \frac{99N}{100}$ respectively and divide the nondecreasing ordered set of values of the variable into 100 equal parts. Thus Q_2, D_5 and P_{50} are same as median.

To determine the quartiles, deciles and percentiles first determine the class $L - U$ of the variable in terms of L = lower boundary U = upper boundary, in which particular quartile (or decile or percentile) lies.

Then we use the formula

$$Q_i = L + \frac{\frac{iN}{4} - F}{f_{Q_i}} \times c, \; i = 1, 2, 3,$$

$$D_i = L + \frac{\frac{iN}{10} - F}{f_{D_i}} \times c, \; i = 1, 2, \ldots, 9 \text{ and}$$

$$P_i = L + \frac{\frac{iN}{100} - F}{f_{P_i}} \times c, \; i = 1, 2, \ldots, 99$$

where L = lower boundary of the class in which i-th quartile Q_i (or i-th decile D_i or i-th percentile P_i) lies, c = class length of that class,

$f_{Q_i}, f_{D_i}, f_{P_i}$ = frequencies of the classes in which Q_i, D_i and P_i lies,

F = cumulative frequency (less than type) of preceeding class of the class containing Q_i(or D_i or P_i).

3.3 Worked out examples

Example 1. Calculate the arithmetic mean, geometric mean, harmonic mean, median, mode of the following data 100, 108, 144, 164, 144, 186.

Solution : A.M. $= \frac{100+108+144+164+144+186}{6} = \frac{846}{6} = 141.$

Let G.M. = G.

Then $G = \sqrt[6]{100\times108\times144\times164\times144\times186}$

$\log G = \frac{1}{6}[\log 100 + \log 108 + \log 144 + \log 164 + \log 144 + \log 186]$

$= \frac{1}{6}[2 + 2\cdot03342 + 2\times2\cdot15836 + 2\cdot21484 + 2\cdot26951]$

$= \frac{1}{6}[4\cdot03342 + 4\cdot31672 + 4\cdot48435]$

$= \frac{12\cdot83449}{6} = 2\cdot13908$

So G = antilog $2\cdot13908 = 137\cdot7463$

H.M. $= \frac{6}{\frac{1}{100}+\frac{1}{108}+\frac{1}{144}+\frac{1}{164}+\frac{1}{144}+\frac{1}{186}}$

$= \frac{6}{\cdot01+\cdot0093+2\cdot\times0069+\cdot0061+\cdot0054}$

$= \frac{6}{\cdot0193+\cdot0138+\cdot0115} = \frac{6}{\cdot0446}$

$= 134\cdot5291.$

When the values are arranged in nondecreasing order, they are

100, 108, 144, 144, 164, 186.

Then two middlemost values are 144 and 144.

So median $= \frac{144+144}{2} = 144.$

Mode = the value which occur most frequently = 144.

Example 2. The average weight of the following distribution is 59 kg.

Weight (in kg) :	50	55	60	x	70	Total
No. of persons :	1	4	2	2	1	10

Find x. For this derived frequency distribution obtain geometric mean, harmonic mean, median and mode.

Solution : Calculation of average weight

Weight (w)	frequency (f)	fw
50	1	50
55	4	220
60	2	120
x	2	$2x$
70	1	70
Total	10	$460 + 2x$

Now average weight $\bar{w} = \frac{\sum fw}{\sum f} = \frac{460 + 2x}{10}$

Given $\bar{w} = 59$. Thus $\frac{460 + 2x}{10} = 59$

or, $460 + 2x = 590$ or, $2x = 130$ or, $x = 65$

Thus missing value $x = 65$ kg.

To get g.m., h.m., median and mode, we construct the following table.

weight (w)	50	55	60	65	70	Total
Frequency (f)	1	4	2	2	1	$10 = N$
Cumulative freq (less than type)	1	5	7	9	10	
$\log w$	1·69897	1·74036	1·77815	1·81291	1·84510	
$f \log w$	1·69897	6·96144	3·55630	3·62582	1·84510	$17·68813 = \sum f \log w$
f/w	·02	·0727	·0333	·0308	·0143	$0·1717 = \sum f/w$

Let G = Geometric mean. Then

$$\log G = \frac{1}{N}\Sigma f \log w = \frac{17{\cdot}68813}{10} = 1{\cdot}768813 \simeq 1{\cdot}7688$$

So G = antilog $1{\cdot}7688 = 58{\cdot}727$ kg.

Let H = Harmonic mean. Then

$$H = \frac{N}{\Sigma \frac{f}{w}} = \frac{10}{0{\cdot}1717} = 58{\cdot}2411 \text{ kg.}$$

$\frac{N}{2} = 5$, $\frac{N}{2} + 1 = 6$. As total frequency is 10, even number, median is the average of two middlemost values i.e., the average of two weights whose C.F.(less than type)'s are 5 and 6 i.e., average of 55 and 60.

So median $= \frac{55+60}{2} = 57{\cdot}5$ kg.

Mode is the weight whose frequency is maximum i.e., frequency is 4, So Mode = 55 kg.

Example 3. Calculate arithmetic mean, geometric mean, harmonic mean, median, mode, first and third quartile from the following frequency distribution of height (in cm.) of 70 persons.

Height (in cm)	126-135	136-145	146-155	156-165	166-175	176-185
No. of persons	7	10	14	23	12	4

Solution : Here variable is height (in cm) of a person and frequency (f) is number of persons. Let M, G, H, be the arithmetic mean, geometric mean and harmonic mean.

Class of heights (cm)	126-135	136-145	146-155	156-165	166-175	176-186	Total
Class in terms of class boundaries	125·5-135·5	135·5-145·5	145·5-155·5	155·5-165·5	165·5-175·5	175·5-185·5	
Mid point (x)	130·5	140·5	150·5	160·5	170·5	180·5	
frequency (f)	7	10	14	23	12	4	70
cum freq (less than type) (F)	7	17	31	54	66	70	

Class of heights (cm)	126-135	136-145	146-155	156-165	166-175	176-186	Total
$u = \frac{x-150{\cdot}5}{10}$	- 2	- 1	0	1	2	3	
fu	- 14	- 10	0	23	24	12	35
$f \log x$	14·80927	21·47676	30·48551	50·72593	26·78069	9·02591	153·30407
$\frac{f}{x}$	·05364	·07117	·09302	·14330	·07038	·02216	·45367

$$M = a.m. = 150{\cdot}5 + 10 \times \frac{\Sigma fu}{N} = 150{\cdot}5 + 10 \times \frac{35}{70} = 150{\cdot}5 + 5 = 155{\cdot}5 \text{ cm.}$$

$$\log G = \frac{1}{N} \Sigma f \log x = \frac{153{\cdot}30407}{70} = 2{\cdot}19006.$$

So, G = g.m. = antilog 2·19006 = 154·90306 cm.

$$H = \frac{N}{\Sigma \frac{f}{x}} = \frac{70}{0{\cdot}45367} = 154{\cdot}2972 \text{ cm.}$$

Median class is 155·5 – 165·5 and its frequency is $f = 23$, lower boundary of median class is $L = 155{\cdot}5$, cumulative freq of former class to median class is $F = 31$, N = Total freq = 70, class length = 10 cm.

$$\text{Median} = L + \frac{\frac{N}{2} - F}{f_m} \times c = 155{\cdot}5 + \frac{35-31}{23} \times 10 = 155{\cdot}5 + \frac{40}{23} = 155{\cdot}5 + 1{\cdot}74$$

$$= 157{\cdot}24 \text{ cm.}$$

Modal class is 155·5 – 165·5. Its frequency = $f_0 = 23$, frequency of former and later classes are $f_{+1} = 12$ and $f_{-1} = 14$ respectively. L = lower boundary of modal class = 155·5, class length = c = 10 cm.

$$\text{Mode} = L + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times c = 155{\cdot}5 + \frac{23-14}{46-14-12} \times 10 = 155{\cdot}5 + \frac{90}{20}$$

$$= 155{\cdot}5 + 4{\cdot}5 = 160 \text{ cm.}$$

First quartile Q_1 lies in the class 145·5 – 155·5 since $\frac{N}{4} = \frac{70}{4} = 17{\cdot}5$ is the cumulative freq (< type) lies in the said class, N = total freq. cumulative freq. of former class = $F = 17$, f = frequency of this class = 14. Class length c = 10 cm, L = lower boundary of the class = 145·5

$$Q_1 = L + \frac{\frac{N}{4} - F}{f} \times c = 145{\cdot}5 + \frac{17{\cdot}5 - 17}{14} \times 10 = 145{\cdot}5 + \frac{5}{14}$$

$$= 145{\cdot}5 + {\cdot}357 = 145{\cdot}857 \text{ cm.}$$

Similarly Q_3 lies in the class $155{\cdot}5 - 165{\cdot}5$ since $\frac{3N}{4} = 52{\cdot}5 < 54$ but $52{\cdot}5 > 31$. Here $L = 155{\cdot}5$, $F = 31$, $f = 23$, $c = 10$ cm.

$$\text{Thus } Q_3 = L + \frac{\frac{3N}{4} - F}{f} \times c = 155{\cdot}5 + \frac{52{\cdot}5 - 31}{23} \times 10 = 155{\cdot}5 + \frac{215}{23}$$

$$= 155{\cdot}5 + 9{\cdot}348 = 164{\cdot}848 \text{ cm.}$$

Example 4. The mean marks in statistics of 100 examinees in a class was 72. The mean marks of boys was 75 while their number was 70. Find the mean marks of girls in the class.

Solution : Let $\bar{x}_1$ = mean marks of n_1 boys in statistics,

$\bar{x}_2$ = mean marks of n_2 girls in statistics

and $\bar{x}$ = mean marks of all the examines (i.e., $n_1 + n_2$) in statistics in the class.

Then $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$. Given $\bar{x} = 72$, $\bar{x}_1 = 75$,

$n_1 = 70$, $n_1 + n_2 = 100$. So $n_2 = 100 - 70 = 30$

So $72 = \frac{70 \times 75 \times 30 \times \bar{x}_2}{100}$ or, $7200 = 70 \times 75 + 30\bar{x}_2$

or, $30\bar{x}_2 = 7200 - 5250 = 1950$

or, $\bar{x}_2 = \frac{1950}{30} = 65$

So ,the mean marks of the girls in statistics of the class is 65.

Example 5. Mean of 100 items is found to be 30. If at the time of calculation, two items are wrongly taken as 32 and 12 instead of correct values 23 and 11. Find the correct mean.

Solution : Wrong total of 100 observations = $30 \times 100 = 3000$

Correct total of 100 observations $= 3000-(32+12)+(23+11)$

$= 3000 - 44 + 34 = 2990$

So correct mean $= \frac{\text{correct total}}{100} = \frac{2990}{100} = 29{\cdot}9$

Example 6. An aeroplane flies around a square the sides of which measure 100 kms each. The aeroplane covers at a speed of 100 kms. per hour the first side, at 200 kms. per hour the second side, at 300 kms. per hour the third side and at 400 kms. per hour the fourth side. Find the average speed and explain the choice of average.

Solution : Here harmonic mean is the appropriate mean since distance covered = speed × time i.e., time $= \frac{\text{distance covered}}{\text{speed}}$ and average speed $= \frac{\text{total distance covered}}{\text{total time taken}} = \frac{d_1 + d_2 + \ldots + d_k}{\frac{d_1}{s_1} + \frac{d_2}{s_2} + \ldots + \frac{d_k}{s_k}}$ where time is unknown but k distances $d_1, d_2, \ldots, d_k$ are covered in k speeds $s_1, s_2, \ldots, s_k$. But if time and speeds are known i.e., in k times $t_1, t_2, \ldots, t_k$ distances are covered with speeds $s_1, s_2, \ldots, s_k$ then average speed $= \frac{\text{total distance covered}}{\text{total time taken}} = \frac{s_1t_1 + s_2t_2 + \ldots + s_kt_k}{t_1 + t_2 + \ldots + t_k}$ which is weighted arithmetic mean of k speeds.

Reqd average speed $= \frac{100+100+100+100}{\frac{100}{100}+\frac{100}{200}+\frac{100}{300}+\frac{100}{400}} = \frac{400}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}}$

$= \frac{400}{\frac{12+6+4+3}{12}} = \frac{4800}{25} = 192$ kms. per hour.

Example 7. A machine depreciates at the rate of 40% of value in first year, 25% in second year and 10% per annum for the next 3 years. What is the average percentage of depreciation for 5 years?

Solution : Let P be initial value. After first year of depreciation the value of the machine is

$$P\left(1-\frac{40}{100}\right)$$

After second year of depreciation the value of machine is

$$P\left(1-\frac{40}{100}\right)\left(1-\frac{25}{100}\right)$$

Thus after 5 years of depreciation the value of the machine is

$$P\left(1-\frac{40}{100}\right)\left(1-\frac{25}{100}\right)\left(1-\frac{10}{100}\right)^3$$

Let $r\%$ be the average percentage of depreciation in 5 years then value after 5 year would be

$$P\left(1-\frac{r}{100}\right)^5$$

Thus $P\left(1-\frac{r}{100}\right)^5 = P\left(1-\frac{40}{100}\right)\left(1-\frac{25}{100}\right)\left(1-\frac{10}{100}\right)^3$

or, $P\left(1-\frac{r}{100}\right)^5 = \cdot 6 \times \cdot 75 \times (\cdot 9)^3 = 0\cdot 45 \times 0\cdot 729$

or, $P\left(1-\frac{r}{100}\right)^5 = 0\cdot 32805$

Taking logarithms, $5 \log\left(1-\frac{r}{100}\right) = \log 0\cdot 32805 = \bar{1}\cdot 5160$

or, $5 \log\left(1-\frac{r}{100}\right) = -.0\cdot 484$

or, $\log\left(1-\frac{r}{100}\right) = -0\cdot 0968 = \bar{1}\cdot 9032$

or, $1-\frac{r}{100} = \text{antilog } \bar{1}\cdot 9032 = \cdot 8002$

or, $\frac{r}{100} = \cdot 1998$ or, $r = 19\cdot 98 \simeq 20$ (approximately)

Hence average rate of depreciation is 20%.

Example 8. Find the missing values from the given information. The median and modal income of 230 persons are Rs. 43·50 and Rs. 44 and the distribution is—

Class of income in Rs.	No. of persons
10-20	4
20-30	16
30-40	—
40-50	—
50-60	—
60-70	8
70-80	2

Solution : The no. of persons = frequency.

Let the frequencies of the classes 30 – 40 be x and 40 – 50 be y. Then frequency of the class 50 – 60 is $230 - 4 - 16 - 8 - 2 - x - y = 230 - 30 - x - y = 200 - x - y$.

Then the cumulative frequency distribution (less than type) [in short, *C. F.*(< type)] is

Class	frequency	C.F. (< type)
10-20	4	4
20-30	16	20
30-40	x	$20 + x$
40-50	y	$20 + x + y$
50-60	$200 - x - y$	220
60-70	8	228
70-80	2	230

Since median is Rs. 43·50, median class is 40 – 50, since 43·50 lies in the class 40 – 50. Then

L = lower boundary of median class = 40

C = class length of median class = 10

f = frequency of median class = y

F = cumulative frequency of preceding class to median class
$= 20 + x$

N = total frequency = 230

Then median $= L + \frac{\frac{N}{2} - F}{f} \times c$

or, $43\cdot50 = 40 + \frac{115-(20+x)}{y} \times 10$

or, $3\cdot50 = \frac{10(95-x)}{y} = \frac{950-10x}{y}$

or, $950 = 10x + 3\cdot5y$ ..... (1)

Further, as mode is Rs. 44 it lies in modal class 40 – 50 and all class lengths are equal.

L = lower boundary of modal class = 40
C = class length of modal class = 10
f_0 = frequency of modal class = y
f_{-1} = frequency of proceding class to modal class = x
f_1 = frequency of proceding class to modal class = $200 - x - y$

Then mode $= L + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times c$

or, $44 = 40 + \frac{y-x}{2y-x-(200-x-y)} \times 10$

or, $4 = \frac{10y-10x}{2y-x-200+x+y}$

or, $4 = \frac{10y-10x}{3y-200}$ or, $12y - 800 = 10y - 10x$

or, $2y + 10x = 800$ ..... (2)

Subtracting (2) from (1) $1\cdot5y = 150$

or, $y = \frac{150}{1\cdot5} = \frac{1500}{15} = 100$

Then from (1) $10x + 350 = 950$ or, $10x = 600$

or, $x = 60$.

So, frequency of the class, 50 – 60 is $200 - x - y = 200 - 160 = 40$.

Thus the required frequency distribution is

Class	No. of persons
10-20	4
20-30	16
30-40	60
40-50	100
50-60	40
60-70	8
70-80	2

3.4 Summary

This chapter consists of central tendency of data, their different measures along with formulae, important properties, their merits and demerits and some useful results. Illustrated examples are given to use these measures and their formulae.

3.5 Exercises

1. Explain what is meant by the central tendency of data. What are the common measures of central tendency?

2. Define arithmetic mean, geometric mean, harmonic mean, median and mode and discuss their merits and demerits as measures of central tendency.

3. Prove that for n real positive observations A.M. $\geq$ G.M. $\geq$ H.M.

4. Prove that for two real positive observations G.M. $= \sqrt{A.M. \times H.M.}$

5. Compare mean, median and mode as measures of central tendency.

6. Show that the combined arithmetic mean $\bar{x}$ of two groups lies between with means $\bar{x}_1$ and $\bar{x}_2$ of two groups.

7. If a variable assumes n values a, ar, ar^2, ..., ar^{n-1} with $r < 1$ and with equal frequencies then find the arithmetic mean (A), geometric mean (G) and harmonic mean (H) and show that $AH = G^2$.

8. Show that sum of deviations of the observations about their mean is zero.

9. Find A.M., G.M., H.M., median and Mode of the numbers (i) 4, 6, 8, 8, 12, 72, (ii) 4, 6, 6, 12, 16.

10. Find the A.M., G.M., H.M., Median, Mode, First and third quartiles of the following frequency distribution.

Daily wages (Rs) :	10	15	20	25	30	35
No. of workers :	5	12	16	14	10	2

11. Find A.M., G.M., H.M., Median, Mode, First and third quariles, 6th decide, 52nd percentile of the following distribution.

Marks :	0-10	10-20	20-30	30-40	40-50
No. of students :	10	20	35	25	10

12. From the following cumulative frequency distribution of marks of 22 students in Mathematics, calculate arithmetic mean, median, mode, first quartile and the number of students whose marks lie between 35 and 45.

Marks :	Below 10	Below 20	Below 30	Below 40	Below 50
No. of students :	3	8	17	20	22

13. The frequency distribution of expenditure of 1000 families is given below.

Expenditure (Rs) :	40-59	60-79	80-99	100-119	120-135
No. of families :	50	?	500	?	50

The mean and median of the distribution are same and equal to Rs. 87·50. Determine the missing frequencies.

14. A company manufacturing 10 KVA generators has divided one of its leading sales territories into 3 zones : urban, semiurban and rural. Ten salesmen working in the urban zone sold 450 units during the year 2002. Another

ten working in semiurban zone sold 510 units and other 20 working in rural zone sold 2000 units in the same period. Find the average number of generators sold per salesman in the whole territory during the period under consideration.

15. A motor car covered a distance of 50 miles four times. The first time at 50 m.p.h., the second at 20 m.p.h., the third at 40 m.p.h., and the fourth at 25 m.p.h., Calculate the average speed.

16. Mr. Sinha take a trip from a plain town to a hill station 60 miles distance at a milage rate of 10 miles per gallon of petrol and on a return trip at 15 miles per gallon. Find the average rate of milage per gallon of petrol for the whole trip.

17. The population of India in 1951 and in 1961 were 361 and 439 millions respectively. (i) What was the average percentage increase per year during the period? (ii) If the average rate of incuase from 1961 to 1971 remains the same what would be the population in 1971?

18. A man gets three successive annual increments in salary of 20%, 30% and 25%, each percentage being reckoned on his salary at the end of the previous year. How much better or worse off would he have been if he had been given 3 annual increments of 25% each, reckoned in same way.

19. In a frequency table the upper boundary of each class interval has a constant ratio to the lower boundary for k classes. Show that geometric mean G may be expressed by the formula.

$$\log G = x_0 + \frac{c}{N}\sum_{i=1}^{k} f_i(i - 1)$$

where x_0 is the logarithm of themid value of the first interval, c the logarithm of the ratio between upper and lower boundaries, f_i = frequency of i-th class, $i = 1, 2, \ldots, k$, N = total frequency.

3.6 Suggest readings

1. Gun, A.M.; Gupta, M.K. and Dasgupta, B. Fundamentals of statistics Vol. I, World Press Pvt. Ltd., 2002.

2. Chaudhuri, S. B. Elementary Statistics Vol. I, Shraddha Prakashani, 1986.

3. Sarkhel, J. and Dutta, Santosh, An insight into Statistics Vol. I, Book syndicate Pvt. Ltd., 1997.

4. Mills, F. G. Statistical Methods, H. Halt, 1955.

5. Yule G.U. and Kendall, M.G. Introduction to the Theory of Statistics, Charles Griffin, 1953.

Unit 4 ❑ Dispersion

Structure

4.0 Objectives

Comparison only in terms of central values, tells a part of the story about two sets of data. This is because how best a particular measure of central value represents the data to which it belongs also depends on the extent of spread of individual observations about the central value.

Consideration of dispersion is necessary because any two sets of data may have the same value of the measure of central tendency. But they may differ

in terms of extent of spread from central value. Thus this spread characteristic of data must be studied and measured.

4.1 Introduction

All the elements of a data set show a tendency to cluster around a specific value, called central value. This gives a general idea of the whole set of data. If the total number of observations and units of measurements of two or more sets of data are same then the measures of central tendency compares the two or more sets of data. But these measure do not compare completely the different sets of data. For this another characteristic, the extent of variability or spread of individual observations from this central value are considered. This characteristic is also called dispersion. Measures of dispersion and their properties are also considered in this chapter.

4.2 Dispersion and its measures

Generally in a set of values the observations are not all equal. But they have a central value arround which they are scattered. This characteristic of spread of different observations of a data-set arround the central value is called dispersion. Dispersion has two types of measures— (i) absolute measures and (ii) relative measures.

Absolute measures of dispersion are Range, Mean deviation, standard deviation, quartile deviation. Relative measures of dispersion are coefficient of variation, coefficient of mean deviation and coefficient of quartile deviation.

4.2.1 Criteria of a good measure

For comparing different measures of this characteristic, dispersion of data, we choose the best one among all the measures, which satisfy the following criteria—

(i) It should be easily understood.

(ii) It should be simple to compute.

(iii) It should be based on all the observations.

(iv) It should not be unduly affected by extreme values.

(v) It should be rigidly defined.

(vi) It should be capable of further algebraic treatments.

(vii) It should not be affected by sampling fluctuations.

4.2.2 Range

It is the difference of lowest value from highest value in a set of observations.

This measure is the simplest possible measure of dispersion to understand and to compute. It takes minimum time to calculate. This measure is not based on all observations. This measure fails to tell this characteristic of the distribution within two extreme observations. It cannot be computed in case of open end classes either in the beginning or at the end. Range of the set of observations 10, 12, 20, 8, 4 is 20 – 4 = 16.

Important property : If $y = a + bx$ be the relation between two variables x and y then range $(y) = |\, b \,|$ range (x) where a and b are constants.

Proof : Maximum $y = a + b$ (*maximum* x) if $b > 0$

$= a + b$ (*minimum* x) if $b < 0$

Minimum $y = a + b$ (*minimum* x) if $b > 0$

$= a + b$ (*maximum* x) if $b < 0$

So if $b > 0$, range (y) = maximum y – minimum y

$= a + b$ (*maximum* x) $- a - b$ (*minimum* x)

$= b$ (*maximum* x – minimum x)

$= b$ (*range* of x)

If $b < 0$, range (y) = maximum y – minimum y

$= a + b$ (*minimum* x) $- a - b$ (*maximum* x)

$= - b$ (*maximum* x – minimum x)

$= - b$ (*range* of x)

So, range $(y) = |\, b \,|$ range (x).

This property shows that the range is independent of change of origin, but depends on the change of scale.