UNIT I

CHAPTER 1

Introduction to Statistics & Sampling Theory

Syllabus

Statistics : Introduction, Origin and Development of Statistics, Definition, Importance and Scope, Limitations, Distrust of Statistics.

Population and Sample : Sampling –Introduction, Types of Sampling, Purposive Sampling, Random Sampling, Simple Sampling, Stratified Sampling, Parameter and Statistic, Sampling Distribution, Statistical Inference, Sampling With and Without Replacement, Random Samples: Random Numbers, Population Parameters, Sample Statistics, Sampling Distributions.

1.1 DEVELOPMENT OF STATISTICS

Sir R. A. Fisher mentioned statistics as : "The science of statistics is essentially a branch of applied mathematics and may be regarded as mathematics applied to **observation** data. Many of the statisticians have given their views as follows.

1. **Lovitt**

   Statistics is the science which deals with the collecting, classifying, presenting, comparing and interpreting numerical data collected to throw light on any sphere of enquiry.

2. **A. L. Bowley**

   (i) Statistics is the device for abbreviating and classifying the statements and making clear the relations.

   (ii) Statistics is the science of measurement of social phenomena regarded as a whole in its manifestations.

   (iii) Statistics is the numerical statement of facts in any department of enquiry, placed in relation to each other.

3. **W. A. Wollis and H. V. Roberts**

   (i) Statistics is not a body of substantive knowledge, but a body of methods obtaining knowledge.

   (ii) Statistics is a body of method for making wise decision in the face of uncertainty.

4. C. H. Meyers

(i) Statistics may be defined as a science of numerical information which employs the processes of measurement and collection, classification, analysis, decision-making and communication of results in a manner understandable and verifiable by other.

(ii) A statistician is a bractitioner of the art and science of statistics.

5. One unknown statisticion said, 'statistics' is the science of averages.' Jokingly he further added ; 'If head is kept in boiler and legs in freeze,' then the temperature of the stomach is statistics.

1.2 ORIGIN AND DEVELOPMENT OF STATISTICS

1) The subject of statistics is as old as the human society itself. In the old age statistics was regarded as the 'Science' of statecraft' and was the by-product of the administrative activity of the state.

2) In the ancient times the scope of statistics was limited to the collection of the following data by the governments for framing military and fiscal policies.

(i) Age and sex-wise population of the country.

(ii) Property and wealth of the country, the former gave the idea of the manner to the government, in order to safeguard itself against any outside aggregation and the latter provided it with the information for the introduction of new taxes and levies.

(3) One of the earliest censuses of population and wealth was conducted by Emperors of Egypt in connection with the construction of famous 'pyramids' Such censuses were later held in England, Germany and other western countries in the middle ages.

(4) In India, an efficient system of collecting official and administrative statistics existed during the reign of Chandragupta Maurya (324-300 B.C.) In Kautilya's 'Arthashastra' a very good system of collecting vital statistics and registration of births and deaths prevailed.

(5) In Germany, the systematic collection of official statistics originated towards the end of 18^{th} century in order to have an idea of the relative strength of different German states, information regarding population and output – industrial and agriculture- was collected.

(6) In England, Napoleonic wars necessitated the systematic collection of numerical data to enable the government to assess the revenues and expenditure with greater precision and then levy the new taxes to meet the cost of war.

(7) During sixteenth century statistics was used for the collection of data to the movements of heavenly bodies-stars and planets-to know about their positions and for the prediction of eclipses.

(8) Tycko Brave (1554-1601) collected the detailed information regarding the movements of the planets, and his student J-Kepler formulated his famous three laws relating to the movements of heavenly bodies. And these laws pared the way for the discovery of Newton's Law of gravitation.

(9) During seventeenth century, **Vital** Statistics was developed by captain John Graunt of London. He made a systematic study of the birth and death statistics.

The computation of mortality tables and the calculation of expectation of life at different ages led to the idea of **'Life Insurance.'** and life insurance Institution was founded in Landon in 1698.

1.2.1 Theory of Probability

(1) The backbone of the theory of statistics is the **'Theory of** probability' or **'The theory of Games and Chance'** was developed in the seventeenth century.

(2) The theory of probability is the outcome of gambling among the nobles of England and France. It was the outcome of estimating the chances of winning or losing in the gamble. The chief contributors were the mathematicians and gamblers of France. Germany and England.

(3) Two French mathematicians Pascal and Fermat solved the Famous 'Problem of Points' posed by the French gambler chevalier de-Mere and this led the foundation stone of the science of probability.

(4) The famous 'Law of Large Numbers' was developed by J. Bernoulli. De Moivre also contributed in this field and published his famous 'Doctrine of chance' he also discovered the normal probability curve which is one of the most important contributions in statistics.

(5) Pierre Simon de Laplace also contributed his monumental work **'Theorie of probability'**. Gauss gave the principle of **Least squares** and established 'the normal Law of Errors.

(6) Quetlet discovered the principle of 'Constancy of Great Numbers' and it formed the basis of sampling **theory.**

Euler, Lagrange, Bayes etc have made outstanding contributions to the modern theory of probability.

(7) Of Russian mathematicians have also mad remarkable contributions to the modem theory of probability. The main contributors to mention only a few of them are ; chebychev, who founded Russian school of statisticians; Markov (Markov chains, Liapoun off (Central limit theorem). A Khinchine (Law of Large numbers), A. Kolmogorov (Who developed calculus of probability), and so on.

(8) To mention modern stalwarts in the development of statistics : Francis Galton pioneered the study of 'Regression Analysis' in Biometry, Karl Pearson pioneered the study of ''Correlation Analysis'. His chi-square test (X^2 -test) of Goodness of Fit is the first and most important of the tests of significance in statistics.

W. S. Cosset with his t-test gave exact sample test.

(9) Most of the work in the statistical theory during the few decades can be attributed to sir Ronald A. Fisher. He applied statistics to a variety of diversified fields such as genetics, biometry, psychology, and educations, agriculture etc. He is termed as Father of statistics. He not only enhanced the existing statistical theory, he pioneered.

(10) **Estimation Theory** (Point Estimation and Fiducial Inference). Exact (small) sampling distribution, Analysis of variance and Design of Experiments.

One writer has described his contributions to the , subject of statistics as :

"R. A. Fisher is the real giant in the development of the theory of statistics."

Indian statisticians also have made significant contributions to the development of statistics in various fields.

(11) The valuable contributions of C. R. Rao (Statistical Inference), Parthasarthy (Theory of probability); P. C. Mahalanobis and P. V. Sukhatme (Sample Surveys); S. N. Roy (Multivariate Analysis); R.C. Bose, K. R. Nair, J. N. Srivastava (Design of Experiments), have placed India's name in the world map of statistics.

1.3 DEFINITION OF STATISTICS

(1) The field of utility of statistics is increasing steadily and thus different people defined it differently according to the development of the subject.

(2) Today statistics covers every sphere of natural and human activity. Hence the old definitions which were confined to a very limited and narrow field of enquiry were placed by the new definitions, which are elaborate in approach.

(3) The word statistics is used to convey different meanings in singular and plural sense.

When used as plural, **statistics means numerical set of data** and when used in singular sense it **means the science of statistical methods embodying** the **theory** and techniques used for collecting, analysing and **drawing inferences from the numerical data**.

We mention below some selected definitions.

(I) Statistics are the classified facts. Representing the conditions of the people in a state specially those facts which can be stated in number or in number or in tables of numbers or in any tabular or classified arrangement'. **– Webster**

(II) 'Statistics are numerical statements of facts in any department of enquiry placed in relation to each other'. **– Bowley.**

(III) 'By statistics we mean quantitative data affected to a marked extent by multiplicity of causes." **– Yule and Kendall.**

(IV) Statistics may be defined as the aggregate of facts affected to a market extent by multiplicity of causes, numerically expressed', enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose, and place in relation to each other'.

– Prof Horace secrist.

Secrist's definition is the most exhaustive of all the four. We examine it in detail.

1.3.1 Analysis of Secrist's Definition of Statistics

(I) Aggregate of Facts

Simple or isolated terms cannot be termed as statistics. They should be a part of aggregate of facts relating to any particular field of enquiry. For example, the height of an individual does not form statistics, but aggregate of the figures of births, deaths, sales, etc. over different times constitute statistics.

(II) Affected by Multiplicity of causes

- Numerical figures should be affected by multiplicity of factors. For example, in social sciences when the effect of some of the factors cannot be measured quantitatively.
- But statistical techniques have been developed to study the joint effect of a number of factors on a single item (Multiple correlation) or the isolated effect of a single factor on the given item (Partial correlation) provided the effect of each of the factors can be measured quantitatively.

(III) Numerically Expressed

- Only numerical data constitute statistics statements like 'the standard of living of the people in Delhi has improved' or 'the production of the particular commodity is increasing' do not contribute statistics.
- In particular, the qualitative characteristics which cannot be measured quantitatively such as intelligence, beauty, honesty cannot be termed as statistics unless they are numerically expressed by assigning particular scores as quantitative standards.
- For example, intelligence is not statistics but intelligence quotient' which is the quantitative measure of the intelligence is statistics.

(IV) Estimated according to reasonable standard of Accuracy

- The numerical data can be obtained by completely enumerating the underlying population
- But, if complete enumeration of the underlying population is not possible (e. g. If population is not infinite, or if testing is destructive etc.) then the data are estimated by using the powerful techniques of sampling and Estimation theory.
- But the estimated values may not be as precise and accurate as the actual values. The degree of accuracy of the estimated values depends on the nature and purpose of enquiry. Here certain standards of accuracy is to be maintained for drawing meaningful conclusions.

(V) Collected in a systematic manner

- The data must be collected in a systematic manner. Thus, for any socio-economic survey, a proper schedule depending upon the object of enquiry should be prepared and trained investigators should be used to collect the data by interviewing the persons.

- The data collected in a haphazard [illegible] and the [illegible] misleading [illegible]

(VI) Collected for a pre-determined purpose

- The objectives of the purpose of the [illegible] should be defined in clear and concrete words [illegible] information [illegible]
- Those commodities or items which are [illegible] belonging to this group should be selected.

(VII) Comparable

- For statistical analysis, data should be comparable. They may be compared with respect to some unit, say time (period) or place.
- For example, the data relating to the population of a country for different years constitute statistics. But the data relating to the [illegible] of an individual and his intelligence quotient (I.Q.) do not constitute statistics since they are not comparable.

1.4 IMPORTNACE AND SCOPE OF STATISTICS

(1) The importance of statistics is explained in the following words by [illegible] Wright (U.S. Commissioner of the Bureau of Labour).

(2) "To a very striking degree our culture has become a statistical culture. Even a person who may never have heard of an index number is affected [illegible] of those index numbers which describe the cost of living."

(3) It is impossible to understand psychology, Sociology, Economics, Finance or a physical science without some general idea of the meaning of an average, of variation, of concomitance, of sampling, of how to interpret charts and table.

(4) According to H.G. Wells: "Statistical thinking will one day be as necessary for effective citizenship as the ability to read and write."

"A knowledge of statistics is like a knowledge of foreign language or of algebra, it may prove of use at any time under any circumstance."

Now, we discuss the importance of statistics in some different disciplines.

1.4.1 Statistics in Planning

(1) The modern age is called as the 'age of planning' and almost all organisations [illegible] the government or businesses or management [illegible] resorting to planning [illegible] efficient working and for formulating policy decisions.

(2) To achieve this, the statistical data relating to production [illegible] investment, income, expenditure and [illegible] and the advanced [illegible] techniques such as index numbers, time series analysis, [illegible] forecasting techniques for handling such data are [illegible].

(3) Today efficient planning is a must for the developing [illegible] economic development.

- The data collected in a hapazard way will not conform to the reasonable standards of accuracy and the conclusions based on them may lead to misleading decisions.

(VI) Collected for a pre-determined purpose

- The objectives or the purpose of the enquiry must be defined in clear and concrete terms. Irrelevant information of data need not be selected.
- Those commodities or items which are consumed or utilised by persons belonging to this group should be selected.

(VII) Comparable

- For statistical analysis, data should be comparable. They may be compared with respect to some unit, say time (period) or place.
- For example, the data relating to the population of a country for different years constitute statistics. But the data relating to the size of the shoe of an individual and his intelligent quotient (I. Q.) do not constitute statistics since they are not comparable.

1.4 IMPORTNACE AND SCOPE OF STATISTICS

(1) The importance of statistics is explained in the following words by Carrol D. Wright, (U. S. Commissioner of the Bureau of Labour).

(2) "To a very striking degree our culture has become a statistical culture. Even a person who may never have heard of an index number is affected ...by... of those index numbers which describe the cost of living;

(3) It is impossible to understand psychology, Sociology, Economics, Finance or a physical science without some general idea of the meaning of an average of variation, of concomitance, of sampling, of how to interpret charts and table."

(4) According to H-6. Wells : "Statistical thinking will one day be as necessary for effective citizenship as the ability to read and write."

"A knowledge of statistics is like a knowledge of foreign language or of algebra : it may prove of use at any time under any circumstances."

Now, we discuss the importance of statistics in some different disciplines.

1.4.1 Statistics in Planning

(1) The modern age is called as the 'age of planning' and almost all organisations in the government or businesses or management are resorting to planning for efficient working and for formulating policy decisions.

(2) To achieve this, the statistical data relating to production, consumption, prices, investment, income, expenditure and so on, and the advanced statistical techniques such as index numbers, time series analysis, demand analysis and forecasting techniques for handling such data are important.

(3) Today efficient planning is a must for the developing economies for their economic development.

(4) In order that planning is successful, it must be based on correct and sound analysis of complex statistical data.

(5) The national sample survey (N. S. S.) was primarily set up in 1950 for the collection statistical data for planning in India.

1.4.2 Statistics in State

(1) With the concept of the idea of welfare state statistical data relating to prices, production, consumption, income and expenditure, investments and profits etc. are used by the governments in formulating economic policies.

(2) Statistical data and techniques are indispensable to the government for planning future economic programms.

1.4.3 Statistics in Mathematics

(1) Statistics is dependent upon mathematics the modem theory of statistics has its foundations on the theory of probability which in turn is a particular branch of mathematical theory of measure and integration.

(2) The development of statistical techniques and theories for application to various science social, physical and natural are based on fitting different mathematical models to the observed data under certain assumption is basically mathematical in character. And a wide application of mathematical tools of integration, differentiation, algebra, trigonometry, matrix theory and so on.

1.4.4 Statistics in Economics

(1) Statistical data and advanced techniques of statistical analysis have proved useful in the solution of variety of economic problems such as production, consumption, distribution of income and wealth, wages, prices, profits, savings, expenditure, investment, unemployment, poverty etc.

(2) For example, the studies of consumption statistics reveal the pattern of the consumption of the various commodities by different sections of the society and also enable us to have some idea about their purchasing capacity and their standard of living.

1.4.5 Statistics in Business and Management

(1) After the industrial revolution, the developments in business activities have taken huge dimensions both in size and the competition in the market. The activities of many of the business enterprises are confined not only to one particular locality, but to larger areas.

(2) Some of the leading houses have the network of their business activities in many of the leading towns and cities of the country, and abroad.

(3) Here the statistical data and the powerful statistical tools of probability, expectation, sampling techniques, tests of significance, estimation theory, forecasting techniques play an important role.

(4) According to wallis and Roberts : "Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty"

(5) Prof. Y-Lun-Chou Further added : " Statistics is a method of decision making i the face of uncertainty on the basis of numerical data and calculated risks."

(6) Business forecasting techniques which are based on the compilation of usefu statistical information are very useful for obtaining estimates which serve as guide to future economic events.

(7) The **time series analysis** is a very important statistical tool which is used i business for the study of :

(i) Trend to obtain the estimates of the probable demand of the goods; trend can be obtained by the method of curve fitting by the principle of leas squares.

(ii) To determine the 'Business cycle', i.e. seasonal and cyclical movements i the phenomenon, and it is composed of prosperity (period of boom) recession, depression and recovery.

(8) The upswings and downswings in business depend on the cumulative nature o the economic forces and the interaction between them.

(9) The studies of **Economic Barometers** (index numbers of prices) enable the businessman to have an idea about the purchasing power of money. The statistical tools of demand analysis enable the businessman to strike a balance between supply and demand.

(10) Statistical techniques are also used by business organisations in

(i) Carrying out **Time and Motion** studies (Which are a part of scientific management).

(ii) **Marketing Decisions** (based on the statistical analysis of consume preference studies – demand analysis).

(iii) **Investment** based on sound study of individual shares and debentures.

(iv) **Personal administration** for the study of statistical data relating to wages cost of living, incentive plans, effect of labour dispute/unrest on the production, performance standards, etc.)

(v) **Credit policy**

(vi) **Inventory control** (for co-ordinations between production and sales)

(vii) **Accounting** (for the evaluation of assets of the business s concerns).

(viii) **Sales control** (through the statistical data pertaining to market studies consumer preference studies, trade channel studies and readership surveys etc.)

"Without an adequate understanding of the statistical methods, the investigato may be like the blind man groping in a dark room for a black cat that is not there.

The methods of statistics are useful in an over-widening range of human activities in any field of though in which numerical data may be had "

– **Croxton and Cowde**

(5) Prof. Y-Lun-Chou Further added : " Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks."

(6) Business forecasting techniques which are based on the compilation of useful statistical information are very useful for obtaining estimates which serve as a guide to future economic events.

(7) The **time series analysis** is a very important statistical tool which is used in business for the study of :

(i) Trend to obtain the estimates of the probable demand of the goods; trend can be obtained by the method of curve fitting by the principle of least squares.

(ii) To determine the 'Business cycle', i.e. seasonal and cyclical movements in the phenomenon, and it is composed of prosperity (period of boom), recession, depression and recovery.

(8) The upswings and downswings in business depend on the cumulative nature of the economic forces and the interaction between them.

(9) The studies of **Economic Barometers** (index numbers of prices) enable the businessman to have an idea about the purchasing power of money. The statistical tools of demand analysis enable the businessman to strike a balance between supply and demand.

(10) Statistical techniques are also used by business organisations in

(i) Carrying out **Time and Motion** studies (Which are a part of scientific management).

(ii) **Marketing Decisions** (based on the statistical analysis of consumer preference studies – demand analysis).

(iii) **Investment** based on sound study of individual shares and debentures.

(iv) **Personal administration** for the study of statistical data relating to wages, cost of living, incentive plans, effect of labour dispute/unrest on the production, performance standards, etc.)

(v) **Credit policy**

(vi) **Inventory control** (for co-ordinations between production and sales)

(vii) **Accounting** (for the evaluation of assets of the business s concerns).

(viii) **Sales control** (through the statistical data pertaining to market studies, consumer preference studies, trade channel studies and readership surveys, etc.)

"Without an adequate understanding of the statistical methods, the investigator may be like the blind man groping in a dark room for a black cat that is not there.

The methods of statistics are useful in an over-widening range of human activities in any field of though in which numerical data may be had. "

– Croxton and Cowden.

1.5 LIMITATIONS OF STATISTICS

Statistics plays dominant role in almost all sciences – social, physical, and natural and is widely used in all spheres of human activity even then it is not without limitations which restrict its scope and utility:

1.5.1 Statistics does not Study Qualitative Phenomenon

(1). Since statistics is a science of dealing with a set of numerical data, it can be applied to the study of those phenomena which can be measured quantitatively.

(2) Thus the statements like 'standard of living of the people in Pune has gone up as compared with last year,' 'population of India has increased considerable during last few years,' do not constitute statistics.

(3) Also, statistics cannot be used directly for the study of quality characteristics like health, beauty, honesty, welfare, poverty etc., and these cannot be measured quantitatively.

1.5.2 Statistics does not Study Individuals

(1) According to prof. Secrist, "By statistics we mean aggregate of facts affected to a marked extent by multiplicity of factors ... and placed in relation to each other."

(2) Hence a single or isolated Figure cannot be regarded as statistics.
[But if it is a part of the aggregate of facts relating to any particular field of enquiry, then it is not regarded as isolated Figure.]

(3) For example, the price of a single commodity, the profit of a particular concern or the production of a particular business house do not constitute statistics, since these figures are unrelated. This is a serious limitation of statistics.

(4) Hence statistics is confined to only those problems where group characteristics are to be studied.

1.5.3 Statistical Laws are Not Exact

(1) Since the statistical laws are probabilistic in nature, inferences based n them are only approximate and not exact. These inferences are not based on mathematical or scientific laws.

(2) Statistical laws are true only on the average.

(3) If the probability of getting head in a single throw of a coin is $\frac{1}{2}$, it does not imply that if we toss a coin 10 times, we shall get 5 heads and 5 tails. But if the experiment of throwing the coin is carried out indefinitely, the we can expect 50% heads and 50% tails.

1.5.4 Statistics is Liable to be Misused

(1) Statistics must be used by experts. According to Bowley, "Statistics only furnishes a tool though imperfect which is dangerous in the hands of those who do not know its use and deficiencies."

(2) The greatest limitations of statistics is that it deals with Figures which are innocent in them selves and do not bear on their face the label of their quality and hence can be easily distorted, manipulated and moulded by dishonest or unskilled workers, unscrupulous people for personal selfish motives.

(3) Statistics neither proves nor disproves anything. It is merely a tool, which if rightly used may prove useful but if misused by inexperienced, unskilled statisticians might lead to fallacious results and conclusions.

In the words of W.I. King, "Statistics are like clay of which you can make a God or a Devil as you please." And,

"Science of statistics is the useful servant but only of great value to those who understand its proper use.

1.6 DISTRUST OF STATISITICS

Irresponsible, inexperienced, and dishonest person who use statistical data and statistical techniques to fulfil their selfish motives have discredited the science of statistics with some very interesting comments, some of which are :

(i) An ounce of truth will produce tons of statistics.

(ii) Statistics can prove anything.

(iii) Figures do not lie. Liars Figure.

(iv) Statistics is an unreliable science.

(v) There are three types of lies-lies, damned lies and statistics, wilked in the order of their naming.

Hence, if statistics and its tools are misused, the fault, does not lie with the science of statistics, Rather, it is the people who misuse it, are to be blamed.

Utmost care and precautions should be taken for the interpretation of statistical data in all its manifestations.

1.7 POPULATION AND SAMPLE

1.7.1 Universe or Population

(1) In a statistical investigation, the interest lies in studying the various characteristics relating to items or individuals belonging to a particular group. This group of individuals under study is known as the **population or universe.**

(2) For example, if we want to study the quality of the manufactured product in an industrial concern during the day, then the population will consist of the day's total production. Thus, "In statistics, population is the aggregate of objects, animate or inanimate, Under study in any statistical investigation. (W. I. King.)

(3) In sampling theory, the population means the larger group from which the samples are drawn.

(i) A population containing a finite number of objects or items is known as **finite population,** e.g. the students in a college.

(ii) A population having number of objects so large as to appear practically infinite, is called as **infinite population**.

1.7.2 Classification of Population

The population may be classified as **existent** or **hypothetical**.

(i) A population consisting of concrete objects is known as **existent population**. e.g. the population of **the books in a library**.

(ii) If the population consists of imaginary objects then it is called the **hypothetical population.**

For example, the population of the throws of a die or a coin, thrown infinite number of times are hypothetical populations.

1.7.3 Sampling

(1) A finite subset of the population selected from it to investigate the properties of the population is called a **sample.**

(2) The number of units in the sample is known as the **sample size.**

(3) Sampling is a tool which helps us to draw conclusions about the characteristics of the population after studying only those objects or items included in the sample.

(4) The main objectives of the sampling theory are :

(i) To obtain maximum information about the characteristics of the population with the available sources in terms of time, money, manpower. It is to be done by studying the sample values only.

(ii) To obtain the best possible estimates of the population parameters.

(5) The error involved in approximations about the population characteristics on the basis of the sample is known as **sampling error**.

(6) Sampling is quite often used in our day-to-day practical life. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

1.7.4 Types of Sampling

The choice of an appropriate sampling design is of great importance in a sample survey. And it is made keeping in view the objective and scope of the enquire and the type of the universe to be sampled.

The sampling techniques may be classified as follows :

1.	Purposive sampling.	2.	Random sampling.
3.	Simple sampling.	4.	Stratified sampling.
5.	Parameter and statistic.	6.	Systematic sampling
7.	Multistage sampling	8.	Quota sampling.

We discuss these types :

1. Purposive sampling

- Purposive sampling is one in which the sample units are selected with definite purpose in view. For example, if we want to give the picture that the standard of living is increased in the Pune city, we may consider individuals in the sample from rich and posh localities, like Deccan Gymkhana, Sindh Colony, Navi Peth

1.7.2 Classification of Population

The population may be classified as **existent** or **hypothetical.**

(i) A population consisting of concrete objects is known as **existent population**, e.g. the population of **the books in a library.**

(ii) If the population consists of imaginary objects then it is called the **hypothetical population.**

For example, the population of the throws of a die or a coin, thrown infinite number of times are hypothetical populations.

1.7.3 Sampling

(1) A finite subset of the population selected from it to investigate the properties of the population is called a **sample.**

(2) The number of units in the sample is known as the **sample size.**

(3) Sampling is a tool which helps us to draw conclusions about the characteristics of the population after studying only those objects or items included in the sample.

(4) The main objectives of the sampling theory are :

(i) To obtain maximum information about the characteristics of the population with the available sources in terms of time, money, manpower. It is to be done by studying the sample values only.

(ii) To obtain the best possible estimates of the population parameters.

(5) The error involved in approximations about the population characteristics on the basis of the sample is known as **sampling error.**

(6) Sampling is quite often used in our day-to-day practical life. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

1.7.4 Types of Sampling

The choice of an appropriate sampling design is of great importance in a sample survey. And it is made keeping in view the objective and scope of the enquire and the type of the universe to be sampled.

The sampling techniques may be classified as follows :

1.	Purposive sampling.	2.	Random sampling.
3.	Simple sampling.	4.	Stratified sampling,
5.	Parameter and statistic.	6.	Systematic sampling
7.	Multistage sampling	8.	Quota sampling.

We discuss these types :

1. Purposive sampling

- Purposive sampling is one in which the sample units are selected with definite purpose in view. For example, if we want to give the picture that the standard of living is increased in the Pune city, we may consider individuals in the sample from rich and posh localities, like Deccan Gymkhana, Sindh Colony, Navi Peth,

Prabhat Road and ignore the localities where low income group and middle class families live.

This sampling suffers from the drawback of favouritism and nepotism and does not give a representative sample of the population.

2. Random Sampling

- A random sample is one in which each unit of population has an equal chance of being included in it.

  In this case the sample units are selected at random and the drawback of favouritism is completely overcome.
- Let us suppose that we take a sample of size r from a finite population of size n. Then there are ^nC_r possible samples.
- A sampling technique in which each of the ^nC_r samples has an equal chance of being selected is known as **random sampling** and the sample obtained by this technique is a **random sample**.
- Proper care has to be taken to ensure that the selected sample is random. Fairly good random samples can be obtained by the use of **Tippet's random number tables** or by throwing of a die, draw of a lottery, etc.
- Normally the method used is **lottery system**, we illustrate the lottery method.
- Suppose we want to select 'r' candidates out of n. We assign numbers one to n, one number (1 to n) on n slips. These slips are made homogeneous in shape, size etc. These slips are put in a bag and shuffled thoroughly and then 'r' slips are drawn one by one. The 'r' candidates corresponding to the numbers on the slips drawn, will constitute the random sample.

Note : Tippet's Random Numbers

Tippet's random number tables consist of 10400 four-digit numbers, giving in all 10400 × 4 = 41600 digits. These are taken from British census reports.

These tables have proved to be fairly random in character. Any page of the table is selected at random and the numbers in any row or column or diagonal selected at random may be taken to constitute the sample.

3. Simple sampling

- Simple sampling is random sampling in which each unit of the population has an equal chance, say p, of being included in the sample and that this probability is independent of the previous drawings.
- Thus a simple sample of size n from a population may be identified with series of n independent trials with constant probability 'P' of success for each trial.

Remark

- Simple sampling implies random sampling but random sampling does not imply simple sampling.
- For example, if an urn contains 'a' white balls and 'b' black balls, the probability of drawing a white ball at the first draw is, say, $\frac{a}{a+b} = p_1$, and if the ball is not replaced, the probability of getting a white ball in the second draw is $\frac{(a-1)}{(a+b-1)} = p_2$ and $p_2 \neq p_1$, hence the sampling is not simple. But since in the first draw each white ball has the same chance

Prabhat Road and ignore the localities where low income group and middle class families live.

This sampling suffers from the drawback of favouritism and nepotism and does not give a representative sample of the population.

2. Random Sampling

- A random sample is one in which each unit of population has an equal chance of being included in it.

  In this case the sample units are selected at random and the drawback of favouritism is completely overcome.
- Let us suppose that we take a sample of size r from a finite population of size n. Then there are ${}^{n}C_{r}$ possible samples.
- A sampling technique in which each of the ${}^{n}C_{r}$ samples has an equal chance of being selected is known as **random sampling** and the sample obtained by this technique is a **random sample.**
- Proper care has to be taken to ensure that the selected sample is random. Fairly good random samples can be obtained by the use of **Tippet's random number tables** or by throwing of a die, draw of a lottery, etc.
- Normally the method used is **lottery system,** we illustrate the lottery method.
- Suppose we want to select 'r' candidates out of n. We assign numbers one to n, one number (1 to n) on s slips. These slips are made homogeneous in shape, size etc. These slips are put in a bog and shuffled thoroughly and then 'r' slips are drawn one by one. The 'r' candidates corresponding to the numbers on the slips drawn, will constitute the random sample.

Note : Tippet's Random Numbers.

Tippet's random number tables consist of 10400 four-digit numbers, giving in all $10400 \times 4 = 41600$ digits. These are taken from British census reports. These tables have proved to be fairly random in character. Any page of the table is selected at random and the numbers in any row or column or diagonal selected at random may be taken to constitute the sample.

3. Simple sampling

- Simple sampling is random sampling in which each unit of the population has an equal chance, say p, of being included in the sample and that this probability is independent of the previous drawings.
- Thus a simple sample of size n from a population may be identified with series of n independent trials with constant probability 'P' of success for each trial.

Remark

- Simple sampling implies random sampling but random sampling does not imply simple sampling.
- For example, if an urn contains 'a' white balls and 'b' black balls, the probability of drawing a white ball at the first draw is, say, $\frac{a}{a+b} = p_1$, and if the ball is not replaced, the probability of getting a white ball in the second draw is $\frac{(a-1)}{(a+b-1)} = p_2$ and $p_2 \neq p_1$, hence the sampling is not simple. But since in the first draw each white ball has the same chance,

i.e., $\left(\frac{a}{a+b}\right)$ of being drawn and in the second draw again each white ball has the same chance,

i. e., $\frac{(a-1)}{(a+b-1)}$ of being drawn. Hence the sampling random.

- Thus, in this case the sampling is random but not simple.
- But if the population is finite and if the sampling is done with replacement, then it is simple sampling.

4. Stratified sampling

- In this type of sampling, the entire heterogeneous population is divided into a number of homogeneous groups, and they are called as strata, they differ from are another but each group is homogeneous within itself.
  Then units are sampled at random from each of these stratum : the sample size in each stratum varies according to the relative importance of the stratum in the population.
- The sample, which is the aggregate of the sampled units of each of the stratum, is termed as **stratified sample.** And the technique of drawing this sample is called as stratified sampling.
- Such a sample is by for the best and can safely be considered as representative of the population from which it has been drawn.

5. Parameter and statistic

- The statistical constants of the population, i.e. mean (μ), variance (σ^2), standard deviation (σ) are referred to as parameters, statistical measures computed from the sample observations alone, e.g., mean $\bar{x}$, Variance (s^2), etc. have been termed as statistics.
- In practice parameter values are not known and estimates based on the sample values are used.
- Thus, statistic which may be regarded as an estimate of parameter, obtained from the sample is a function of the sample values only.
- We observe that a statistic, as it is based on sample values and as there are multiple choices of the samples that can be drawn from a population, varies from sample to samples.
- The determination of the variation in the values of the statistic obtained from different samples. May be attributed to chance or fluctuations of sampling, is one of the fundamental problems of the sampling theory.

Remarks

(1) We note that, μ and σ^2 will refer to the population mean and variance respectively while the sample mean and variance will be denoted by $\bar{x}$ and s^2 respectively.

(2) **Unbiased Estimate :** A statistic $t = t(x_1, x_2, \ldots, x_n)$, a function of the sample values $x_1, x_2, \ldots, x_n$ is an unbiased estimate of the population parameter θ, if $e(t) = \theta$.
In other words, if E (statistic) = Parameter, then statistic is said to be an unbiased estimate of the parameter.

6. Systematic sampling

- Systematic sampling is slight variation of the simple random sampling in which only the first sample unit is selected at random and the remaining units are automatically selected in a definite sequence of equal spacing from one another. This technique of drawing samples is usually recommended if the units are arranged in some systematic order such as alphabetical, chronological, geographical order, etc.
- This requires the sampling units in the population to be ordered in such a way that each item in the population is uniquely identified by its order, for example the list of voters, the name of the persons in a telephone directory etc.

Let us suppose that N sampling units in the population are arranged in some systematic order and serially numbered from 1 to N and we want to draw a sample of size n from it such that

$N = nk, \quad \therefore K = \frac{N}{n}$, where k is called as the **sample interval**.

- Systematic sampling consists in selecting any unit at random from the first k units numbered from 1 to k and then selecting every k unit in succession.

If the first unit selected at random is i unit, then sample of size n will consist of the units :

$i, i + k, i + 2k, \ldots, i + (n - 1) k$

- The number i is called as the random start.

For example, suppose that we want to select 100 voters from a list of 1,000 names, then here n = 100, N = 1000, $\therefore K = \frac{N}{n} = \frac{1000}{100} = 10$.

- Now, we select any number from 1 to 10 at random and the corresponding voter in that list is selected.
- Suppose the selected number is 4, then the sample will consist of voters containing 1000 names as :

4, 14, 24, ..., 994.

We can select k possible systematic samples starting with 1, 2 ,....k unit as :

Random start	Sample composition (units in the sample)					
1	1	1 + k	.....	1 + jk	...	1 + (n – 1) k
2	2	2 + k	.....	2 + jk	...	2 + (n – 1) k
:	:	i + k	.....	:	...	:
:	:	:	.....	i + jk	...	i + (n – 1) k :
k	k	2k	.....	(1 + j)k	...	: nk

Merits of systematic sampling

(i) Systematic sapling is easy to operate and checking can be done quickly. It results in saving of time and labour relative to simple random sampling.

(ii) It is more efficient than simple random sampling provided the units are arranged serially

Demerits of systematic sampling

(i) It works well if the **complete** frame is available and units are arranged serially. But these requirements are generally not fulfilled.

(ii) Systematic sampling gives biased results if there are periodic features in the frame and the sampling interval (K) is equal to or a multiple of the period.

Cluster sampling

Here the total population is divided into some recognisable sub divisions. Which are termed as **clusters** and a simple random sample of these cluster is drawn.

In using cluster sampling, the following points should be kept in mind :

(i) Clusters should be as small as possible consistent with the cost and limitations of survey.

(ii) The number of sampling units in each cluster should be approximately same.

7. Multistage Sampling

- Here sub-sampling within the clusters is carried out. This technique is called two-stage sampling, clusters being termed as primary units and the units within the clusters as secondary units.

This technique can be generalised to **multi-stage sampling.**

Multi-stage sampling consists in sampling first stage units by some suitable method of sampling.

Then a sub-sample is drawn by some suitable method of sampling. Then further stages are added to arrive at a sample of desired sampling units.

Merits

(i) Multistage sampling is more flexible as compared to other methods of sampling.

(ii) It is simple to carry out and becomes administratively convenient as it covers large area.

(iii) Its practical advantage is that we need the second stage only for those units which are selected in the first stage sample and it saves operational cost.

Demerits

(i) The variability of the estimates under this methods is greater than that of estimation based on simple random sampling.

(ii) The variability depends upon the composition of the primary units.

(iii) A multistage sampling is usually less efficient than a suitable single stage sampling of the same size.

8. Quota sampling : Quota sampling is a special form of stratified sampling.

- In this method, an investigator examines the number of sample units from the stratum assigned to him. The quota of units to be examined is fixed. It is fixed depending on specified characteristics such as income group, sex, occupation, political or religious affiliations, etc.
- The investigator applies his judgement in the choice of the sample and tries to get the desired information.

6. Systematic sampling

- Systematic sampling is slight variation of the simple random sampling in which only the first sample unit is selected at random and the remaining units are automatically selected in a definite sequence of equal spacing from one another. This technique of drawing samples is usually recommended if the units are arranged in some systematic order such as alphabetical, chronological, geographical order, etc.
- This requires the sampling units in the population to be ordered in such a way that each item in the population is uniquely identified by its order, for example the list of voters, the name of the persons in a telephone directory etc.

Let us suppose that N sampling units in the population are arranged in some systematic order and serially numbered from 1 to N and we want to draw a sample of size n from it such that.

$N = nk, \therefore K = \frac{N}{n}$, where k is called as the **sample interval.**

- Systematic sampling consists in selecting any unit at random from the first k units numbered from 1 to k and then selecting every k unit in succession.

If the first unit selected at random is i unit, then sample of size n will consist of the units :

i, i + k, i + 2k, ... i + (n – 1) k.

- The number i is called as the random start.

For example, suppose that we want to select 100 voters from a list of 1,000 names, then here n = 100, N = 1000, $\therefore K = \frac{N}{n} = \frac{1000}{100} = 10$.

- Now, we select any number from 1 to 10 at random and the corresponding voter in that list is selected.
- Suppose the selected number is 4, then the sample will consist of voters containing 1000 names as :

4, 14, 24, ..., 994.

We can select k possible systematic samples starting with 1, 2 ,....k unit as :

Random start	Sample composition (units in the sample)					
1	1	1 + k	.....	1 + jk	...	1 + (n – 1) k
2	2	2 + k	.....	2 + jk	...	2 + (n – 1) k
:	:	i + k	.....	:	...	⋮
:	:	:	.....	i + jk	...	i + (n – 1) k :
k	k	2k	.....	(1 + j)k	...	⋮ nk

Merits of systematic sampling

(i) Systematic sapling is easy to operate and checking can be done quickly. It results in saving of time and labour relative to simple random sampling.

(ii) It is more efficient than simple random sampling provided the units are arranged serially

Demerits of systematic sampling

(i) It works well if the **complete** frame is available and units are arranged serially. But these requirements are generally not fulfilled.

(ii) Systematic sampling gives biased results if there are periodic features in the frame and the sampling interval (K) is equal to or a multiple of the period.

Cluster sampling

Here the total population is divided into some recognisable sub-divisions. Which are termed as **clusters** and a simple random sample of these cluster is drawn.

In using cluster sampling, the following points should be kept in mind :

(i) Clusters should be as small as possible consistent with the cost and limitations of survey.

(ii) The number of sampling units in each cluster should be approximately same.

7. Multistage Sampling

- Here sub-sampling within the clusters is carried out. This technique is called two-stage sampling, clusters being termed as primary units and the units within the clusters as secondary units.

This technique can be generalised to **multi-stage sampling.**

Multi-stage sampling consists in sampling first stage units by some suitable method of sampling.

Then a sub-sample is drawn by some suitable method of sampling. Then further stages are added to arrive at a sample of desired sampling units.

Merits

(i) Multistage sampling is more flexible as compared to other methods of sampling.

(ii) It is simple to carry out and becomes administratively convenient as it covers large area.

(iii) Its practical advantage is that we need the second stage only for those units which are selected in the first stage sample and it saves operational cost.

Demerits

(i) The variability of the estimates under this methods is greater than that of estimation based on simple random sampling

(ii) The variability depends upon the composition of the primary units.

(iii) A multistage sampling is usually less efficient than a suitable single stage sampling of the same size.

8. Quota sampling : Quota sampling is a special form of stratified sampling.

- In this method, an investigator examines the number of sample units from the stratum assigned to him. The quota of units to be examined is fixed. It is fixed depending on specified characteristics such as income group, sex, occupation, political or religious affiliations, etc.
- The investigator applies his judgement in the choice of the sample and tries to get the desired information.

In case of non-response from some of the selected sample units, the investigator. Selects some fresh units himself to complete his quota.

☞ **Merits**

(i) Quota sampling is a stratified-cum-purposive or judgement sampling and thus enjoys the benefits of both.

(ii) If tries at making the best use of stratification without incurring high costs involve in any probabilistic method of sampling.

(iii) There is considerable saving in time and money as the sampled units may be so selected that they are close together.

(iv) If carefully executed by skilled and experienced investigators, quota sampling gives quite reliable results.

☞ **Demerits**

(i) Since quota sampling is a restricted type of judgement sampling, it suffers from all the limitations of judgement or purposive sampling, i.e.,

(ii) It may be biased because of personal beliefs and prejudices of the investigator in the selection of the units or/and inspecting them.

(iii) It may involve bias due to the substitution of the sampled units from where there is no response.

(iv) Since it is not based on random sampling, the sampling error cannot be estimated.

In spite of that, the technique of quota sampling, is generally adopted in market surveys, political surveys, or surveys of option poll, where it is difficult to identify the strata in advance.

1.8 SAMPLING DISTRIBUTION

If we draw a sample of size n from a given finite population of size N, then the total number of possible samples is :

$$^{N}C_{n} = \frac{N!}{n!(N-n)!} = k \qquad \text{(say)}$$

For each of these k samples, we can evaluate some statistic $t = t(x_1, x_2, \ldots, x_n)$, in particular the mean $\bar{x}$, the variance s^2, etc. an given below :

Sample number	T(statistic)	Statistic $\bar{x}$	S^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
⋮	⋮	⋮	⋮
k	T_k	$\bar{x}_k$	s_k^2

The set of values of the statistic so obtained, one for each sample, constitutes is called **sampling distribution** of the statistic.

For example, the valves $t_1, t_2, \ldots t_k$ determine the sampling distribution of the statistic t.

In other words, **statistic t may** be regarded as a random variable, which take the values $t_1, t_2, \ldots t_k$ and we can compute the variour statistical constants like mean, variance standard deviation etc. for its distribution.

For example, the mean and variance of the sampling distribution of the statistic 't' are given by.

$$\bar{t} = \frac{1}{k}(t_1 + t_2 + \ldots + t_k) = \frac{1}{k}\sum_{i=1}^{k} t_i,$$

$$\text{and } \text{Var}(t) = \frac{1}{k}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \ldots + (t_k - \bar{t})^2\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k}(t_i - \bar{t})^2$$

1.8.1 Standard Error

- The standard deviation of the sampling distribution of a statistic is known as its **Standard Error,** denoted by S.E.
- We mention below the standard errors of some of the well-known statistics, for large samples :

No.	Statistic	Standard error
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Sample Variance : s^2	$\sigma^2\sqrt{2/n}$
3.	Sample S. D. : S	$\sqrt{\sigma^2/2n}$
4.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
5.	Difference of two sample s.d.s. $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$

1.9 STATISTICAL INFERENCE

- Statistical inference is defined as the procedure of analysing the result and making conclusion from data based on random variation.
- Statistical inference is the technique of making decisions a but the parameters of a population that relies on random sampling.
- It enables us to assess the relationship between dependent and independent variables.

- The idea of statistical inference is to estimate the uncertainty or sample to sample variation. It helps use to deliver a range of values for the true value of something in the population.
- The components use for making the statistical inference are :
  (i) Sample size,
  (ii) Variability in the sample
  (iii) Size of the observed difference.

1.9.1 Types of Statistical Inference

There are different types of statistical inference that are used to draw conclusions such as Pearson Correlation, Biraviate Regression, Multivariate Regression, Anova or T-test and chi-square statistic and contingency table.

But, the most important two types of statistical inference that are primarily used are :

(i) Confidence Interval
(ii) Hypothesis testing,

1.9.2 Statistical Inference Procedure

Following steps follow the statistical inference procedure :

(i) Start with a theory,
(ii) Design research hypothesis,
(iii) Implement the variables
(iv) Acknowledge the population to which Results should be applied.
(v) Draw up the null hypothesis for this population.
(vi) Assemble the sample of children from the population and begin the study.
(vii) Implement the statistical test to examine if the collected sample properties are sufficiently different from what is expected under the null hypothesis to be able to reject the null hypothesis.

1.9.3 Statistical Inference Solutions

- Statistical inference solutions deliver efficient use of statistical data with respect to the group of individuals or trials.
- It manages with every character, including the collection, investigation and analysis of data and organising the collection of data.
- Through statistical inference solutions, people can gain knowledge after initiating their work in multiple fields.
- Some of the statistical inference solution facts are as follows :
  (i) It is an usual method to predict that the observed samples are independent observations from a population type such as Poisson or normal.
  (ii) Statistical inference solution helps to evaluate the parameter (s) of the expected model such as normal mean or binomial proportion.

1.9.4 Importance of Statistical Inference

- To make an effective solution, accurate data analysis is important to interpret the results of the research.
- Inferential statistics is used in the future prediction for varied observations in different fields.
- It enables us to make inferences about the data. It also helps us to deliver a probable range of values for the true value of something in the population.

  Statistical inference is used in different fields such as :
  (i) Business analysis, (ii) Artificial intelligence,
  (iii) Financial analysis, (iv) fraud detection,
  (v) Machine learning, (vi) Pharmaceutical Sector,
  (vii) Share Market

1.9.5 Statistical Inference Example

Ex. 1.9.1 : A bag containing 2 yellow balls, 3 red balls and 5 black balls. One ball is drawn at random from the bag what is the probability that the black ball is drawn.

☑ **Soln. :**

Through statistical inference solution

Total number of balls in a bag = 2 + 3 + 5 = 10.

$$\text{Number of black balls} = 5$$

$$\text{probability of black ball} = \frac{\text{No. of black balls}}{\text{Total no. of balls}} = \frac{5}{10} = \frac{1}{2}$$

$$\therefore \text{ probability of black balls} = \frac{1}{2}$$

1.10 SAMPLING WITH AND WITHOUT REPLACEMENT

- If the units are selected or drawn one by one in such a way that a unit drawn at a time is replaced back to the population before the subsequent draw then it is known as simple random sampling with replacement(SRSWR).
- In this type of sampling from a population of size N, the probability of selection of a unit at each draw remains $\frac{1}{N}$.
- In SRSWR, a unit can be included more than once in a sample. Therefore if the required sample size is n, the effective sample size is sometimes less than n due to the inclusion of one or more units more than once.
- With the idea that effective sample size be adhered to the simple random sampling without replacement is adopted. In this method a unit selected once is not included in the population at any subsequent draw. Hence, the probability of drawing a unit from a population of N units at r draw is $\left(\frac{1}{N-r+1}\right)$.

- In simple random sampling, the probability of selection of any sample of size n from a population consisting of N units remains the same, $\frac{1}{\binom{N}{n}}$, i.e. $\binom{N}{n}$ is the number of all possible samples.

1.10.1 Random Numbers

Lottery method, that we have described, is quite time consuming and complicated to use especially when the population to be sampled is sufficiently large. Also, in this method it is not practicable to make all the slips or cards exactly alike and hence some **bias** is likely to be introduced.

To avoid this, statisticians have considered the random sampling **number series.** Most of these series are the results of actual sampling operations recorded for future use.

The most practical and inexpensive method of selecting a random sample consists in the use of 'Random Number Tables.' It is so constructed that each of the digits 0, 1, 2, ..., 9 appears with approximately the same frequency and independently of each other. If we have to select a sample from a population of size N (≤ 99), then the numbers can be combined two by two to give pairs from 00 to 99.

Similarly if $N \leq 999$ or $N \leq 9999$ and so on, then combining the digits three by three (or four by four and so on), we get numbers from 000 to 999 or 0000 to 9999 and so on. Since each of the digits 0, 1, 2, ...9 occurs with approximately the same frequency and independently of each other, so does each of the pairs 00 to 99, triplets 000 to 999 or quadruplets 0000 to 9999 and so on.

The method of drawing a random sample comprises the following steps :

(i) Identify N units in the population with the number 1 to N.

(ii) Select at random, any page of the **'random number table'** and pick up the numbers in any row, column or diagonal at random.

(iii) The population units corresponding to the numbers selected in step (ii) constitute the random sample.

There are different sets of random numbers commonly used in practice. The numbers in these tables have been subjected to various statistical tests for randomness of a series and their randomness has been well established for all practical purposes.

1.10.1.1 Solved Examples

Ex. 1.10.1 : The adjoining table of ten random numbers of two digits each is provided to field the investigator.

34	96	61	85	49
78	50	02	27	13

How should he use this table to make a random selection of 5 plots out of 40.

Soln. :

Step (1) : We first identity the 40 plots with the numbers 1 to 40. In the given table there are only 3 numbers, i.e., 34, 02, 13 which are less than 40 and hence it is not possible to draw the desired sample of size 5 from the table.

So, we assign more than 1 number to each of the sampling units, i.e., plots.

For example, the **first plot** will be assigned the numbers.

$01, 01 + 40, 01 + 2 \times 40, ...$

i.e. 1, 41, 81, 121, ...

Similarly the second plot will be assigned the numbers :

$02, 02 + 40, 02 + 2 \times 40, ...$

i.e. 02, 42, 82, 122, ...

Now, the last plot, i.e. 40^{th} plot can be assigned the number.

0, 40, 80, 120, ...

If we select the first number (in the table) and move row-wise, we set, the table

Number from table	Number of the sampled plots
34	34
$96 = 16 + 2 \times 40$	16
$61 = 21 + 40$	21
$85 = 5 + 2 \times 41$	5
$49 = 9 + 40$	9

Thus the plot nos. 5, 9, 16, 21 and 34 constitute the desired sample

1.11 POPULATION PARAMETERS

- In statistics, a parameter is any measured quantity of a statistical population that summarises or describes an aspect of the population, such as a mean or a standard deviation.
- If a population exactly follows a known and defined distribution, for example the **normal distribution,** then a small set of parameters can be measured which completely describes he populations, and it can be considered to define a **probability distribution.** This is for the purposes of extracting **samples** from this populations.
- In short, we can say, a **parameter** is to a **population** as a **statistic** is to a sample
- A parameter describes the **true value** calculated from the full population, where as a statistic is an estimated measurement of the parameters based on a sub sample.
- Hence a 'statistical parameter' is referred to as a **population parameters :**

1.11.1 Types of Parameters

Parameters are given names appropriate to their roles, including the following :

(i) Location parameter (ii) Dispersion parameter or scale parameter

(iii) Shape parameter.

Where a probability distribution has a domain over a set of objects that are themselves probability distribution, the terms **concentration** parameters is used.

Quantities such a **regression coefficients** are statistical parameters in the above sense. It is because they index the family of **conditional probability distributions** that describe how the **dependent variables** are related to the independent variables.

1.11.2 Measurement of Parameters

- In statistical inference, parameters are sometimes taken to be unobservable.
- Estimators of a set of parameters of a specific distribution are often measured for a population, under the assumption that the population is distributed according to that specific probability distribution.
- Even if a family of distribution is not specified, quantities such as the **mean** and **variance** can still be regarded as statistical parameters of the population. And statistical procedures can still attempt to make inferences about such population parameters.

1.12 SAMPLE STATISTICS

- Sampling is the selection of a subset (a statistical sample) of individuals from within a **statistical population** to estimate characteristics of the whole population.
- Sampling has lower costs and faster data collection than measuring the entire population and can provide insights in cases where it is not practicable to measure an entire population.
- Each **observation** measures one or more properties (such as weight, location, colour or mass) of independent objects or individuals.
- In business and medical research, sampling is widely used for gathering information about a population.

1.12.1 Sampling Methods

- A variety of sampling methods can be employed individually or in combination.
- Factors commonly influencing the choice between these designs include.
  - (i) Nature and quality of the frame
  - (ii) Availability of auxiliary information about units on the frame.
  - (iii) Accuracy requirements, and the need to measure accuracy.
  - (iv) Whether detailed analysis of the sample is expected.
  - (v) Cost/operational concerns.

1.12.2 Applications of Sampling

- Sampling enables the selections of right data points from within the larger data set to estimate the characteristics of the whole population
- In manufacturing different types of sensory data such as Acoustics, vibration, pressure, current, voltage and controller data are available at short time intervals.

- To predict down-time. It is not necessary to look at all the data but a sample may be sufficient.

1.12.3 Errors in Sample Surveys

- Survey results are typically subject to some error.
- Total errors can be classified into sampling errors and non-sampling errors.
- The term 'error' implies systematic **biases** as well as random errors.

1.12.3.1 Sampling Errors and Biases

Sampling errors and biases are induced by the sample design. They include

(i) **Selection bias :** When the true selection probabilities differ from those assumed in calculating the results,

(ii) **Random sampling error :** Random variation in the results due to the elements in the samples being selected at random

1.12.3.2 Non-Sampling Error

Non-Sampling errors are other errors which can impact final survey estimates, caused by problems in data collection, processing, or sample design. Such errors may include :

(i) **Over-coverage :** inclusion of data from outside of the population.

(ii) **Under-coverage** : Sampling frame does not include elements in the population.

(iii) **Measurement error :** For examples when respondents misunderstand a question, or find it difficult to answer.

(iv) **Processing error :** Mistakes in data coding.

(v) **Non-response or participation bias :** failure to obtain complete data from all selected individuals.

1.13 SAMPLING DISTRIBUTIONS

- In statistics, a sampling distribution or finite-sample distribution is the probability distributions of a given random-sample based statistic.
- If an arbitrarily large number of samples, each involving multiple observations (data points), were separately used in order to compute one value of a statistics (such as, for example, sample mean or sample variance) for each sample, then the sampling distribution is the probability distribution of the values that the statistic takes on.
- Even if only one sample is observed, the sampling distribution can be found theoretically.
- Sampling distributions are important in statistics because they provide a major simplification to statistical inference.
- More clearly, they allow analytical considerations to be based on the probability distribution of a statistic, rather than on the joint probability distribution of all the individual sample values.

1.13.1 Sampling Distribution of a Statistic

The **sampling distribution** of a statistic is the **distribution** of that statistic, considered as a **random variable**, when derived from a **random sample** of size n. It may be considered as the distribution of the statistic for all **possible samples from the same population** of a given sample size.

- The sampling distribution depends on the underlying, **distribution** of the population; we consider the statistic and employ sampling procedure and use the sample size. For example, consider a normal population with mean μ and variance σ^2.
- We assume that we repeatedly take samples of a given size from this population and calculate the **arithmetic mean** $\bar{x}$ for each sample-the statistic is called the sample mean.
- The distribution of these means, or averages, is called the "Sampling distribution of the sample mean."
- This distribution is normal N (μ, σ^2/n) (n is the sample size), since the given population is normal sampling distributions may often be close to normal even when the population distribution is not normal.

1.13.2 Standard Error

- The standard deviation of the sampling distribution of a **statistic** is referred to as the **standard error** of that quantity. For the case, when the statistic is the sample mean, and samples are uncorrelated, the standard error is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Where σ is the standard deviation of the population distribution and n is the sample size (i.e. number of items in the sample).
- An important result of this formula is that the sample size must be quadrupled (multiplied by 4 to achieve half) $\left(\frac{1}{2}\right)$ the measurement error.

When designing statistical stuolies where cost is a factor, this may have a role in under-standing cost-benefit tradeoffs.

- For the case where the statstic is the sample total, and samples are uncorrelated, the standard error is :

$$\sigma_{\sum x} = \sigma\sqrt{n}$$

Where, again, σ is the standard deviation of the population distribution of that quantity and n is the sample size (number of items in the sample).

1.13.3 Examples

Population	Statistic	Sampling distribution
Normal : N (μ, σ^2)	Sample mean $\bar{x}$ from samples of size n	$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ If the standard deviation σ is not known, one can consider, $T = (\bar{x} - \mu)\frac{\sqrt{n}}{S}$, which follows the student's t-distribution with $\gamma = n - 1$ degrees of freedom. Here S^2 is the sample variance, and T is a **pivotal quantity**, Whose distribution does not depend on σ
Bernoulli Bernoulli (P)	Sample proportion of 'Successful' trials $\bar{x}$	$n\bar{x} \sim$ Bernoulli (n, p)
Two independent normal populations : N (μ_1, σ_1^2) and N (μ_2, σ_2^2)	Difference between sample means : $\bar{x}_1 - \bar{x}_2$	$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
Any absolutely continuous distribution F with density f	Median $X_{(k)}$ from a sample of size n = 2k − 1, where sample is ordered $X_{(1)}$ to $X_{(n)}$	$F_{x(k)}(x) = \frac{(2k-1)!}{(k-1)!^2} f(x)\,(f(x)\,(1 - F(x)))^{k-1}$
Any distribution with distribution function f	Maximum M = max. X_k from a random sample of size n	$F_{m(x)} + P(M \le x) = \prod P(X_k \le x) = (F(x))^n$

Chapter End...

❑❑❑

UNIT II

CHAPTER 2

Descriptive Statistics : Measures of Central Tendency

Syllabus

Frequency Distributions and Measures of central Tendency : Frequency Distribution, Continuous Frequency Distribution, Graphic Representation of a Frequency Distribution, Histogram, Frequency Polygon, Averages or Measures of Central Tendency or Measures of Location, Requisites for an Ideal Measure of Central Tendency, Arithmetic Mean, Properties of Arithmetic Mean, Merits and Demerits of Arithmetic Mean, Weighted Mean, Median, Merits and Demerits of Median, Mode, Merits and Demerits of Mode, Geometric Mean, Merits and Demerits of Geometric Mean, Harmonic Mean, Merits and Demerits of Harmonic Mean, Selection of an Average.

2.1 FREQUENCY DISTRIBUTION

- When observations, discrete or continuous, on a single characteristics of a large number of individuals are available, then it becomes necessary to condense the data.
- A **frequency table** is constructed by arranging collected data values in ascending order of magnitude with their corresponding frequencies.
- The **frequency** of a particular data value is the number of times the data value occurs. For example, if four students have a score of 80 in mathematics, then the score of 80 is said to have a frequency of 4.

  The frequency of a data value is often represented by f.

2.1.1 Frequency Table

A frequency table is constructed by arranging collected data values in ascending order of magnitude with their corresponding frequency.

Ex. 2.1.1 : The marks awarded for an assignment set for a year 8th class of 20 students were as follows :

6	7	5	7	7	8	7	6	9	7
4	10	6	8	8	9	5	6	4	8

Present this information is a frequency table ;

☑ **Soln. :**

- **Step (I)** : We construct a table with three columns. The first column shows what is being arranged in ascending order (i.e. the marks). The lowest mark is 4. So, start from 4 in the first column
- **Step (II)** : Go through the list of marks. The first mark in the list is 6, so put a tally mark against 6 in the second column.

  The second mark is 7, so put a tally mark against 7 in the second column.

  The third mark is 5, so put a tally mark against 5 in the second column.

  We continue the process until all marks in the list are tallied.
- **Step (III)** : Count the number of tally marks for each mark and write it in third column. The finished table is **as follows** :

Mark	Tally	Frequency
4	II	2
5	II	2
6	IIII	4
7	IIII	5
8	IIII	4
9	II	2
10	I	1

2.1.2 Class Intervals

When the set of data values are spread out, it is difficult to set up a frequency table for every data value as there will be too many rows in the table. So we group the data into **class intervals**, so that we can organise and analyse the data.

The **frequency of a group** (or class interval) is the number of data values that fall in the range specified by that group (or class interval).

Ex. 2.1.2 : The number of calls from motorists per day for roadside service was recorded for the month of January, 2010. The results were as follows :

28	122	217	130	120	86	80	90	120
140	70	40	145	187	113	90	68	174
194	170	100	75	104	97	75	123	100
82	109	120	81					

Set up a frequency table for this set of data values.

☑ **Soln. :**

We proceed as follows to construct a frequency table :

Smallest data value = 28, Highest data value = 217

Difference = highest value – smallest value = 217 – 28 = 189

Let the width of the class interval be 4

$\therefore$ Number of class intervals $= \frac{189}{40} = 4.7 = 5$ (round up the figure)

∴ There are at least 5 class intervals.

- **Step (1) :** We construct a table with three columns, and then write class-interval in the first column. The size of each group is 40. So, the groups will start at 0, 40, 80, 120, 160 and 200.
- **Step (2) :** We go through the list of data values. For the first data value in the list, 28, place a tally mark against the group 0 – 39 in the second column. For 122, place a tally mark against the group 120- 159 in the second column. We continue the process until all of the data values in the set are tallied.
- **Step (3) :** count the number of tally marks for each group and write it in the third column. The finished frequency table is as shown.

Class interval	Tally	Frequency
0 – 39	1	1
40 – 79	1𝍸1	5
80 – 119	1𝍸1 1𝍸1 11	12
120 – 159	1𝍸1 111	8
160 – 199	1111	4
200 – 239	1	1
	Sum =	31

2.1.3 Continuous Frequency Distribution

When we deal with a continuous variable, it is not possible to arrange the data in the class interval of above type. Let us consider the distribution of age in years. If the class interval is 15-19, 20 – 24, etc; then the persons between 19 to 20 years are not taken into account.

In such a case, we form the class interval as shown :

Age (in years)	→	
5 or more but less than 10		0 – 5
10 or more but less than 15		5 – 10
15 or more but less than 20		10 – 20
and 50 on		15 – 20
		20.....

2.2 HISTOGRAM

Histogram is commonly used device for charting **continuous** frequency distribution. It consists in erecting a series of adjacent vertical rectangles on the section of the horizontal axis (X - axis), with bases (sections) equal to the width of the corresponding class intervals and heights are so taken that the areas of rectangles are equals to the frequencies of the corresponding classes.

2.2.1 Construction of Histogram

The variate values are taken along X-axis and the frequencies along Y-axis.

Case (I) Histogram with equal classes

- If classes are of equal magnitude, each class interval is drawn on X-axis by a section which is equal to the magnitude of the class interval.
- On each class interval erect a rectangle with the height proportional to the corresponding frequency of the class. The series of adjacent rectangles (one for each class) so formed gives the histogram of the frequency distribution and its area represents the total frequency of the distribution.

Case (ii) : Histogram with unequal classes

- If the classes are not uniform, then the different classes are represented on X-axis by sections which are equal to the magnitude of the corresponding classes and the heights of the corresponding rectangles are to be adjusted so that the area of the rectangle is equal to the frequency of the corresponding class.
- This can be done by taking the height of each rectangle equal to the corresponding frequency density of each class, where.

$$\text{Frequency density of a class} = \frac{\text{Frequency of the class}}{\text{Magnitude of the class}}$$

2.2.2 Example

Ex. 2.2.1 : Represent the adjoining distribution of marks of 100 students in the examination by a histogram.

Marks	Obtained	Number of students
Less than	10	4
Less than	20	6
Less than	30	24
Less than	40	46
Less than	50	67
Less than	60	86
Less than	70	96
Less than	80	99
Less than	90	100

☑ **Soln. :** First we convert the given cumulative frequency distribution into the frequency distribution of marks :

Marks	Number of students
0-10	4
10-20	6 – 4 = 2
20-30	24 – 6 = 18
30-40	46 – 24 = 22
40-50	67 – 46 = 21
50-60	86 – 67 = 19
60-70	96 – 86 = 10
70-80	99 – 96 = 3
80-90	100 – 99 = 1

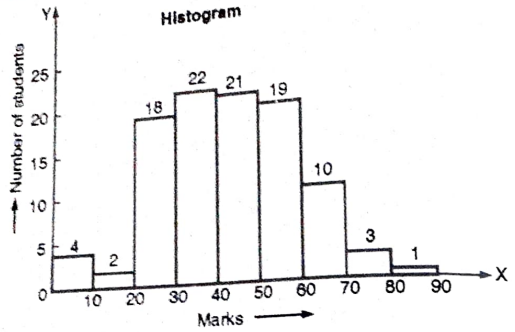

Fig. P. 2.2.1

Ex. 2.2.2 : Represent the following data by means of a histogram :

Weekly wages (100 Rs.)	10-15	15-20	20-25	25-30	30-40	40-60	60-80
Number of workers	7	19	27	15	12	12	8

☑ **Soln. :**

- Here the class-intervals are of unequal magnitude, the corresponding frequencies have to be adjusted to obtain the 'frequency-density', so that area of the rectangle is equal to the class frequency.
- We note that the first four classes are of magnitude 5, the class 30-40 is of magnitude 10 and the last two classes 40-60 and 60-80 are of magnitude 20.
- Since 5 is the minimum class interval, the frequency of the class 30-40 is divided by 2 and the frequencies of classes 40-60 and 60-80 are to be divided by 4 as shown :

Weekly wages (100Rs.)	Number of workers	Magnitude of class	Height of rectangle
10-15	7	5	7
15-20	19	5	19
20-25	27	5	27
25-30	15	5	15
30-40	12	10	$\frac{12}{2} = 6$
40-60	12	20	$\frac{12}{4} = 3$
60-40	8	20	$\frac{8}{4} = 2$

☞ **Histogram**

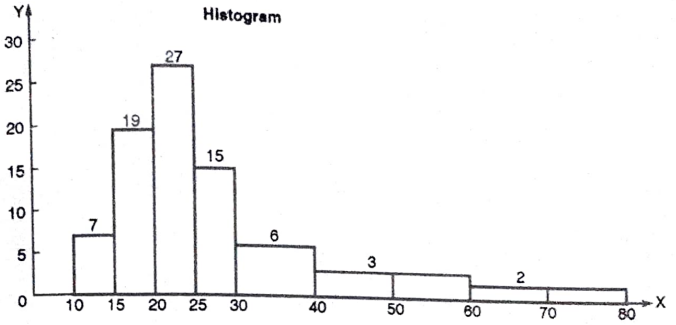

Fig. P. 2.2.2

2.3 PIE-DIAGRAM

Just as rectangles are used to represent the total magnitude and its various components, the circle may be divided into various sections or segments representing certain proportion or percentage of the various component parts to the total. Such a sub-divided circle diagram is known as **an angular or pie diagram.**

2.3.1 Steps for Construction of Pie-Diagram

(i) Express each of the component values as a percentage of the respective total.

(ii) Since the angle at the centre of the circle is 360°, the total magnitude of the various components is taken to be equal to 360°.
The degrees represented by the various component parts of a given magnitude can be obtained as follows :

$$\text{Degree of any component part} = \frac{\text{component value}}{\text{Total value}} \times 360°$$

(iii) Pie-diagram is also known as **circular diagram**

2.3.2 Example

Draw a pie-diagram to represent the following data of proposed expenditure by a state-govt. for the year 1947-98.

Items	Agriculture and rural development	Industries and urban development	Health and education	Miscellaneous
Proposed expenditure (in million Rs.)	4200	1,500	1000	500

☑ **Soln. : Calculation of Pie-chart**

Items	Proposed expenditure	Angle at the centre
(1)	(2)	(3) $\frac{(2)}{7200} \times 360°$
Agriculture and Rural development	4,200	$\frac{4200}{7200} \times 360° = 210°$
Industries and urban development	1500	$\frac{1500}{7200} \times 360° = 75°$
Health and education	1000	$\frac{1000}{7200} \times 360° = 50°$
Miscellaneous	500	$\frac{500}{7200} \times 360° = 25°$
Total	7200	360°

PIE-diagram representing proposed expenditure

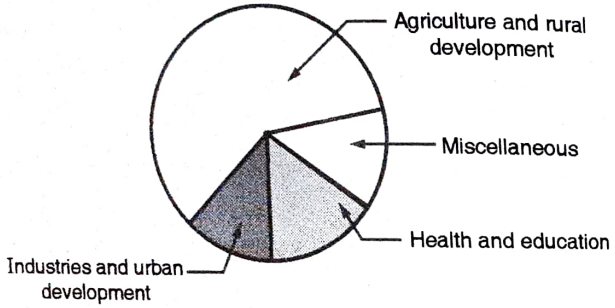

Fig. 2.3.1

2.4 FREQUENCY POLYGON

(i) For an ungrouped distribution, the frequency polygon is obtained by plotting points with abscissa as the variate values and the ordinate as the corresponding frequencies and joining the plotted points by means of straight lines.

(ii) For a grouped frequency distribution, the abscissa of points are mid values of the class intervals.

(iii) For equal class intervals the frequency polygon can be obtained by joining the middle points of the upper sides of the adjacent rectangles of the histogram by means of straight lines.

(iv) If the class intervals are of small width, the polygon can be obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

2.5 AVERAGES (OR MEASURES OF CENTRAL TENDENCY)

- According to Prof. Bowley, averages are "statistical constants which enable us to comprehend in a single effort the significance of the whole".
- An average of a statistical series is the value of the variable which is the representative of the entire distribution.
- The following five measures of central tendency that are common in use. :
  (i) Arithmetic Mean, (ii) Median, (iii) Mode
  (iv) Geometric Mean and (v) Harmonic Mean

2.5.1 Requisites for an Ideal Measure of Central Tendency

According to Prof. Yule, the following are the characteristics to be satisfied by an ideal measure of central tendency :

(i) It should be rigidly defined.

(ii) It should be readily comprehensible and easy to calculate.

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

(iv) It should be suitable for further mathematical treatment. It means that, if we are given the averages and sizes of a number of series, we should be able to calculate the average of the composite series, obtained on combining the given series.

(v) It should be affected as little as possible by fluctuations of the sampling.

(vi) It should not be much affected by extreme values.

2.6 INTRODUCTION

Summarization of data is a necessary function of statistical Analysis. The data summarized in the form of tables and frequency distributions. In order to bring the characteristics of the data, these tables and frequency distributions need to be summarized further. A measure of central tendency or an average is very essential and an important summary measure in any statistical analysis. There are five types of measures of central tendency or average which are comonly used.

(i) Arithmetic mean (ii) Median (iii) Mode
(iv) Geometric mean (v) Harmonic mean

2.6.1 Arithmetic Mean

The arithmetic mean of a set of observations is their sum divided by the number of observation. Let $x_1, x_2, ..., x_n$ be n observations. Then their average or arithmetic mean is given by,

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n}$$

e.g.

The marks obtained by 10 students in class XII in mathematics examinations are 26, 32, 22, 56, 41, 43, 15, 50, 33, 40. The arithmetic mean of the marks is given by,

$$\bar{x} = \frac{\sum x}{n} = \frac{26+32+22+56+41+43+15+50+33+40}{10} = 35.8$$

☞ Arithmetic mean of grouped data

In case of grouped or continuous frequency distribution, the arithmetic mean is given by

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum f_x}{N}, \text{ Where } N = \sum_{i=1}^{n} f_i$$

Here $x_1, x_2, \ldots x_n$ are n observations and $f_1, f_2, \ldots, f_n$ are corresponding frequencies.

☞ Arithmetic mean from assumed mean

If the values of x and f are large the calculation of mean becomes quite time consuming. In such cases the provisional mean 'a' is taken as the value of x (mid-value of the class interval). This number is taken corresponds to highest frequency or near middle value of the frequency distribution R is called assumed mean.

$$\text{Let } d = x - a$$

$$fd = f(x-a) = fx - fa$$

$$\sum fd = \sum fx - a\sum f = \sum fx - aN$$

Dividing both sides by n,

$$\frac{\sum fd}{N} = \frac{\sum fx}{N} - a$$

☞ Arithmetic mean by step deviation method

When the class intervals in a grouped data are equal, calculation can be simplified by step-deviation method. In such cases, deviation of the variate x from the assumed mean a are divided by the common factor h which is equal to the width of the class interval.

$$\text{Let } d = \frac{x-a}{h}$$

$$\bar{x} = a + h\frac{\sum fd}{\sum f} = a + h\frac{\sum fd}{N}$$

Where a is the assumed mean.

$$d = \frac{x-a}{h} \text{ is the deviation}$$

h is the width of the class interval, N is the number of observations

2.6.2 Median

Median is the central value of the variable after arranged in ascending or descending order of magnitude. In case of up grouped data if the number of observations is even, there are two middle terms and the median is obtained by taking arithmetic mean of that two middle terms.

e.g.

(i) The median of the values 20, 8, 23, 27, 13, 15, 30
i.e. 8, 13, 15, 20, 23, 27, 30 is 20, as 20 is the central value.

(ii) The median of the values 7, 21, 31, 51, 25, 15, 30, 41
i.e. 7, 15, 21, 25, 30, 31, 41, 51 is 27.5

☞ Median for continuous frequency distribution

In case of continuous frequency distribution (less than frequency distribution), the class corresponding to the cumulative frequency just greater than $\frac{N}{2}$, is called the median class and the value of the median is given by,

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

Where

l is the lower limit of the median class
f is the frequency of the median class
h is the width of the median class
C is the cumulative frequency of the class proceeding median class.
N is the sum of frequencies.

In case of 'More than' or 'greater than' type of frequency distributions, the value of median is given by,

$$\text{Median} = u - \frac{h}{f}\left(\frac{N}{2} - C\right)$$

Where u is the upper limit of the median class
f is the frequency of the median class
h is the width of the median class
C is the cumulative frequency of the class succeeding the median class.

Note : The class corresponding to cumulative frequency just greater than $\frac{N}{2}$ is called median class.

2.6.3 Mode

Mode is the value which occurs most frequently in a set of observations. In case of discrete frequency distribution, mode is the value of x corresponding to the maximum frequency.

e.g. (i) In the series 6, 7, 9, 3, 5, 8, 9, 5, 3, 4, 5 the value 5 occurs most frequently. Hence mode is 5. (ii) Consider following distribution

x	1	2	3	4	5	6
f	5	8	20	15	9	2

The value of x corresponding to the maximum frequency 20 is 3. Hence mode is 3.

For an asymmetrical frequency distribution the difference between the mean and the mode is approximately three times the difference between the mean and the median.

Mean – mode = 3 (Mean – median); Mode = 3 Median – 2 mean

Which is known as empirical formula for calculation of mode.

☞ Mode for a continuous frequency distribution

$$\text{Mode} = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

Where l is lower limit of modal class

h is the width of the modal class

f_m is the frequency of the modal class

f_1 is the frequency of the class preceding the modal class

f_2 is the frequency of the class succeeding the modal class

➡ **Note :** The class corresponding maximum frequency is called modal class.

2.6.4 Examples on Mean, Median, Mode

REx. 2.6.1 (Ref. - Q. 1(a), S-20, 3 Marks)

Calculate the mean for the following frequency distribution

Class	0-8	8-16	16-24	24-32	32-40
Frequency	8	10	17	12	5

☑ **Soln. :**

Let a = 20 be the assumed mean; d = x = a = x – 20

Class	Frequency	Mid-value (x)	d = x – 20	fd
0-8	8	4	– 16	– 128
8-16	10	12	– 8	– 80
16-24	17	20	0	0
24-32	12	28	8	96
32-40	5	36	16	80
	$\Sigma f = 52$			$\Sigma fd = -32$

$$N = \Sigma F = 52$$

$$\bar{x} = a + \frac{\Sigma fd}{N} = 20 + \frac{(-32)}{52} = 19.3846$$

REx. 2.6.2 (Ref. - Q. 3(c)(i)(OR), S-20, 3 Marks)

The following table gives the distribution of the companies according to size of capital. Find the mean size of the capital of a company also find median capital of a company

Capital (Rs. In lacs)	< 5	< 10	< 15	< 20	< 25	< 30
Number of companies	20	27	29	38	48	53

☑ **Soln. :**

This is a less than type of frequencies distribution. This we have first convert into class interval.

Let a = 12.5 be the assumed mean and h = 5 be the width of the class interval.

Cluss intervals	Cummlutive frequencies (less than)	Frequency (f)	Mid Value (x)	$d = \frac{x - 12.5}{5}$	fd
0-5	20	20	2.5	– 2	– 40
5-10	27	7	7.5	– 1	– 7
10-15	29	2	12.5	0	0
15-20	38	9	17.5	1	9
20-25	48	10	22.5	2	20
25-30	53	5	27.5	3	15
		$\Sigma f = 53$			$\Sigma fd = -3$

$$\text{Mean } \bar{x} = a + h\frac{\Sigma fd}{N} = 12.5 + 5\left(\frac{-3}{53}\right) = 12.22 \text{ lacs}$$

To find median,

$$\text{Since } \frac{N}{2} = \frac{53}{2} = 26.5, \text{ the median class is 5-10.}$$

$$\text{Here } l = 5, h = 5, f = 7, C = 20$$

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - C\right) = 5 + \frac{5}{7}(26.5 - 20) = 9.6429$$

REx. 2.6.3 (Ref. - Q. 1(a), W-20, 3 Marks)

The frequency distribution of marks obtained by 60 students of a class in a college is given by,

Marks	30-34	35-39	40-44	45-49	50-54	55-59	60-64
Frequency	3	5	12	18	14	6	2

Find the mode of the distribution.

☑ **Soln. :**

The class intervals are first converted into continuous exclusive series as shown in the following table.

Marks	Frequency
29.5-34.5	3
34.5-39.5	5
39.5-44.5	12
44.5-49.5	18
49.5-54.5	14
54.5-59.5	6
59.5-64.5	2

Since maximum frequency is 18 which lies in the interval 44.5-49.5 the modal class is 44.5-49.5 Here $l = 44.5$, $h = 5$, $f_m = 18$, $f_1 = 12$, $f_2 = 14$.

$$\text{Mode} = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

$$\text{Mode} = 44.5 + 5\left[\frac{18 - 12}{2(18) - 12 - 14}\right] = 47.5$$

Ex. 2.6.4 : The aggregate of marks obtained by two groups of students are given below. Find out which of the two shows greater variability. Which group is more consistent ?

Group A	518	519	530	530	544	542	518	550	527	527	531	550	550	529	528
Group B	825	830	830	819	814	814	844	842	842	826	832	835	835	840	840

☑ **Soln. :**

To show variability of two group, we have to determine coefficient of variation $\left(\frac{\sigma}{A.M} \times 100\right)$ in each case :

We prepare the data in frequency distribution form :

x	f	d = x – 530	d^2	fd	fd^2
518	2	– 12	144	– 24	288
519	1	– 11	121	– 11	121
527	2	– 3	9	– 6	18
528	1	– 2	4	– 2	4
530	2	0	0	0	0
531	1	1	1	1	1
542	1	12	144	12	144
544	1	14	196	14	196
550	3	20	400	60	1200
	$\Sigma f = 15$			$\Sigma fd = 43$	1973

Now, $$\text{A.M.} = 530 + \left(\frac{\Sigma f.d.}{\Sigma f}\right) = 530 + \frac{43}{15} = 532.866$$

and $$\sigma = \sqrt{\frac{1}{N}\Sigma fd^2 - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{1973}{15} - \left(\frac{43}{15}\right)^2} = 11.105$$

Coefficient of variation $= \frac{\sigma}{A.M.} \times 100$

Coefficient of variation $= \frac{11.105}{532.866} \times 100 = 2.0840$

For group B

x	F	d = x – 830	d^2	f.d.	$f \cdot d^2$
814	2	– 16	256	– 32	512
819	1	– 11	121	– 11	121
825	1	– 5	25	– 5	25
826	1	– 4	16	– 4	16
830	2	0	0	0	0
832	1	2	4	2	4
835	2	5	25	10	50
840	2	10	100	20	200
842	2	12	144	24	288
844	1	14	196	14	196
	$\Sigma f = 15$			$\Sigma fd = 18$	$\Sigma fd^2 = 1412$

Now, $$\text{A.M.} = 830 + \frac{18}{15} = 831.2$$

And $$\sigma = \sqrt{\frac{1412}{15} - \left(\frac{18}{15}\right)^2} = \sqrt{94.133 - 1.44} = 9.628$$

Coefficient of variation $= \frac{\sigma}{A.M} \times 100 = \frac{9.628}{831.2} \times 100 = 1.158$

∴ Coefficient of variation of group A is greater than that of group B.

∴ Group A has greater variability,

i.e.. Group is more consistent.

Ex. 2.6.5 : Calculate the first four moments about the mean of the given distribution. Also find β_1 and β_2

x	2.0	2.5	3.0	3.5	4.0	4.5	5.0
f	4	36	60	90	70	40	10

☑ **Soln. :**

Taking A = 3.5, and $u = \frac{x - 3.5}{0.5}$; we prepare table :

x	f	$u = \frac{x-3.5}{0.5}$	fu	fu^2	fu^3	fu^4
2.0	4	– 3	– 12	36	– 108	342
2.5	36	– 2	– 72	144	– 288	576
3.0	60	– 1	– 60	60	– 60	60
3.5	90	0	0	0	0	0
4.0	70	1	70	70	70	70
4.5	40	2	80	160	320	640
5.0	10	3	30	90	270	810
	$\Sigma f = 310$		$\Sigma fu = 36$	$\Sigma fu^2 = 560$	$\Sigma fu^3 = 204$	$\Sigma fu^4 = 2480$

Now, $\mu_1' = \frac{\Sigma fu}{\Sigma f} = \frac{36}{310} = 0.166$; $\mu_2' = \frac{\Sigma fu^2}{\Sigma f} = \frac{560}{310} = 1.806$

$\mu_3' = \frac{\Sigma fu^3}{\Sigma f} = \frac{204}{310} = 0.658$; $\mu_4' = \frac{\Sigma fu^4}{\Sigma f} = \frac{2480}{310} = 8.0$

Now, $\mu_1 = 0$,

$\mu_2 = \mu_2' - (\mu_1')^2 = 1.806 - 0.013456 = 1.7925$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$

$= 0.658 - 0.6285 + 0.003122 = 0.03262$

and $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

$= 8.0 - 0.3053 + 0.1458 - 0.000543 = 7.8399$

and $\beta_1 = \frac{\mu_3^2}{\mu_2} = \frac{0.001064}{5.7594} = 0.000185$

$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{7.8399}{3.213} = 2.44$

2.7 GEOMETRIC AND HARMONIC MEAN

There are two other averages, the geometric mean and the harmonic mean which are sometimes used.

1. The Geometric Mean (G.M.) of a set of observations is such that its logarithm is equal to the Arithmetic mean of the logarithms of the values of the observations. This is given by the same formula even if the observations occur with certain frequencies.

   Consider the set of values 2, 3, 5, 6 which occur with frequencies 10, 16, 24, 10 respectively. If x is the Geometric Mean then,

   $$\log x = \frac{10\log 2 + 16\log 3 + 24\log 5 + 10\log 6}{10 + 16 + 24 + 10} = 0.5867$$

$\therefore x = 3.86$

2. The harmonic mean (H.M.) of a set of observations is such that its reciprocal is the arithmetic mean (A.M.) of the reciprocals of the values of the observations. Consider the above set of values. The harmonic mean y of the set of values is given by,

   $$\frac{1}{y} = \frac{10\times\frac{1}{2} + 16\times\frac{1}{3} + 24\times\frac{1}{5} + 10\times\frac{1}{6}}{60} = 0.28$$

   $\therefore y = 3.57$

Note :

1. The Geometric mean can be found only if the values assumed by the observations are positive.
2. It can be shown that A.M. ≥ G.M. ≥ H.M.

2.7.1 Other Measures of Location, Quartiles, Deciles and Percentiles

- We have seen that the median of a set of measurements is the value which divides the set into two equal halves, each containing 50% of the measurements.
- In the same way, some other measures of location can be considered. We define the three quartiles, Q_1, Q_2 and Q_3.
- They are such that when the measurements are arranged in increasing order, they divide the set of measurements into four equal parts, the first quartile Q_1 contains the 25% of the measurement, the second quartile Q_2 contains 50% of the measurements and the third quartile Q_3 contains 75% of the measurements.
- Actually the second quartile Q_2 is the median.

  Similarly we define the deciles. The first decile D_1 contains 10% of the measurements, the second decile D_2 contains 20% of the measurements and so on.
- The fifth decile is the median. In the same manner, we define percentiles. The 99 percentiles P_1, ... P_{99} divide the set of measurements into 100 equal parts.
- The first percentile P_1 contains 1% of the measurements, the second percentile P_2 contains 2% of the measurements and so on, the 12^{th} percentile contains 12% of the measurements.
- The 50^{th} percentile is therefore the median. The method of finding out the quartiles, deciles and percentiles is basically the same as that of finding the median.
- The median divides the set of observations into two equal values, each containing 50% of the measurements, the 3^{rd} decile divides the set into two parts, the first part being 30% of the set and the other containing 70% of the observations.

Ex. 2.7.1 : The distribution of fortnightly wages of 280 employees of an undertaking is as given.

Find the first quartile, the median and the third quartile, find D_4, P_{66}, P_{10} and P_{90}.

Table P. 2.7.1

Fortnightly wages (Rs.)	Frequency
Less than 200	12
200-400	16
400-600	38
600-800	78
800-1000	80
1000-1200	35
1200-1400	14
Above 1400	7
Total	**280**

☑ **Soln.:**

First we prepare cumulative frequency Table P. 2.7.1(a).

Table P. 2.7.1(a)

Wages (Rs.)	Frequency	Cumulative Frequency
Less than 200	12	12
200-400	16	28
400-600	38	66
600-800	78	144
800-1000	80	224
1000-1200	35	259
1200-1400	14	273
Above 1400	7	280

▶ **Step I :** To find quartiles

The observation for the first quartile Q_1 corresponds to $\frac{280}{4} = 70^{th}$ observation which lies in the interval 600-800, with lower class boundary 600. This interval contains 78 observations and the interval preceding this contains 66 observations. Hence,

$$Q_1 = l_1 + \frac{l_2 - l_1}{f_1}(m - C)$$

where l_1 = Lower limit of the class in which Q_1 lies

l_2 = The upper limit of HM class in which Q_2 lies

f_1 = Positive frequency of the class

$m = \frac{N}{4}$

C = Cumulative frequency of the group preceding the Q_1 class.

$$\therefore Q_1 = 600 + \frac{200}{78}\left(\frac{280}{4} - 66\right) = 610.25 \text{ Rs.}$$

The median which is the second quartile Q_2, is given by,

$$Q_2 = 600 + \frac{200}{78}\left(\frac{280}{2} - 66\right) = 600 + 189.74 = 789.74 \text{ Rs.}$$

The third quartile Q_3 is given by,

$$Q_3 = 800 + \frac{200}{80}\left[280 \times \frac{3}{4} - 144\right] = 965 \text{ Rs.}$$

▶ **Step II :** The observation for 4^{th} decile corresponds to the $\frac{280 \times 4}{10} = 112^{th}$ observation, which lies in the interval 600 – 800. Hence,

$$D_4 = 600 + \frac{200}{78}(112 - 66) = 717.95 \text{ Rs.}$$

▶ **Step III :** The observation for 66^{th} percentile corresponds to $280 \times \frac{66}{100} = 184^{th}$ observation which lies in the interval 800 – 1000. Thus the 66^{th} percentile P_{66} is given by,

$$P_{66} = 800 + \frac{200}{80}(184.8 - 144) = 902 \text{ Rs.}$$

In the same way,

$$P_{10} = 400 + 0 = 400$$

and $$P_{90} = 1000 + \frac{200}{35}\left[\frac{280 \times 90}{100} - 220\right] = 1182.96 \text{ Rs.}$$

2.8 THE RANGE

The range of set of numbers is the difference between the largest and the smallest items of the set. The range is a very crude measure. It does not tell us about the distribution of the values of the set relative to the average.

2.8.1 The Semi-Interquartile Range

This is a more refined form of range. It is defined by,

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

And is called as semi-interquartile range.

Q_1 = First quartile

Q_3 = Third quartile

2.9 THE MEAN DEVIATION

Consider a set of observations, $x_1, x_2, ..., x_n$. The mean deviation (or average deviation) is defined by,

$$\text{M.D.} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \qquad \text{where,} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

is the arithmetic mean and $|x_i - \bar{x}|$ is the absolute value of the deviation.

Ex. 2.9.1 : Find the mean deviation of the set of measurement 1, 3, 8.

☑ **Soln. :** Here the arithmetic mean,

$$\bar{x} = \frac{1+3+8}{3} = 4$$

$$\therefore \text{M.D.} = \frac{|1-4|+|3-4|+|8-4|}{3} = 2.67$$

2.9.1 Mean Deviation for Grouped Data

Let $x_1, x_2, ..., x_n$ occur with the corresponding frequencies $f_1, f_2, ..., f_n$, then

$$\text{M.D.} = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{\sum_{i=1}^{n} f_i} \qquad \text{where,} \quad \bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

Note that the above formula is also applicable in the case of a frequency distribution whose class intervals have mid-points $x_1, x_2, ..., x_n$ and the classes have frequencies $f_1, f_2, ... f_n$.

Ex. 2.9.2 : Calculate the mean deviation from the mean of the following distribution

Table P. 2.9.2

Marks	Number of students
0-10	5
20-20	8
20-30	15
30-40	16
40-50	6
Total	**50**

☑ **Soln. :**

We first calculate mean then find mean deviation.

Table P. 2.9.2(a)

Mid value	$u = \frac{x-25}{10}$	f	fu	$x - \bar{x}$	$f\|x - \bar{x}\|$
5	– 2	5	– 10	– 22	110
15	– 1	8	– 8	– 12	96
25	0	15	0	– 2	30
35	1	16	16	8	128
45	2	6	12	18	108
Total		**50**	**10**		**472**

Here, Mean $= 25 + \frac{10}{50} \times 10 = 27$ marks and

$$\text{Mean deviation} = \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i} = \frac{472}{50} = 9.44 \text{ marks}$$

Ex. 2.9.3 : The mean annual salary paid to all employees of a company was Rs. 5000. The mean annual salaries paid to male and female employees were Rs. 5200 and Rs. 4200 respectively. Determine the percentage of males and females employed by the company.

☑ **Soln. :** Let n_1 and n_2 represent percentage of males and females respectively.

Then $n_1 + n_2 = 100$ ...(1)

Now, mean annual salary of males $= \overline{x_1} = 5200$ Rs.

Mean annual salary of females $= \overline{x_2} = 4200$ Rs.

Mean annual salary of all employees $= \bar{x} = 5000$ Rs.

$$\text{Now,} \quad \bar{x} = \frac{n_1 \overline{x_1} + n_2 \overline{x_2}}{n_1 + n_2} \quad \text{(Note the formula)}$$

$$\therefore \quad 5000 = \frac{n_1\, 5200 + n_2\, 4200}{100}$$

$\therefore \quad 26\, n_1 + 21\, n_2 = 2500$ ...(2)

From Equations (1) and (2),

$n_1 = 80, n_2 = 20$. $\therefore$ Percentages are 80 and 20.

Ex. 2.9.4 : The first of the two samples has 100 items with mean 15 and standard deviation 3. If the whole group has 250 items with mean 15.6 and standard deviation $\sqrt{13.44}$, find the standard deviation of the second group.

☑ **Soln. :** We have,

$n_1 = 100, \quad \overline{x_1} = 15, \quad \sigma_1 = 3$

$n = n_1 + n_2 = 250$

$\bar{x} = 15.6$ $\quad \therefore n_2 = 250 - 100 = 150$

$\sigma = \sqrt{13.44}$

We use the formula,

$$\bar{x} = \frac{n_1\bar{x_1} + n_2\bar{x_2}}{n_1 + n_2} \quad \therefore 15.6 = \frac{100(15) + 150\,\bar{x_2}}{250}$$

$$\therefore \bar{x_2} = 16$$

The variance of the combined group σ^2 is given by the formula,

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}$$

(Note the formula)

Where $d_1 = \bar{x_1} - \bar{x} = 15 - 15.6 = -0.6$

and $d_2 = \bar{x_2} - \bar{x} = 16 - 15.6 = 0.4$ $\quad \because \sigma = \sqrt{13.44}$

$$\therefore 13.44 = \frac{100(9) + 150\,\sigma_2^2}{250} + \frac{100(-0.6)^2 + 250(0.4)^2}{250}$$

$$\therefore 250 \times 13.44 = 900 + 36 + 40 + 150\,\sigma_2^2 \quad \therefore \sigma_2 = 4$$

Ex. 2.9.5 : Calculate the first four moments of the following distribution about mean and hence find β_1 and β_2.

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

☑ **Soln. :** We first calculate the first four moments about the point x = 4.

We prepare the table.

Table P. 2.9.5

x	f	d = x – 4	fd	fd^2	fd^3	fd^4
0	1	– 4	– 4	16	– 64	256
1	8	– 3	– 24	72	– 216	648
2	28	– 2	– 56	112	– 224	448
3	56	– 1	– 56	56	– 56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
$\sum$	N = 256	–	0	512	0	2816

Recall that,

$\mu_r' = \frac{1}{N}\sum f \cdot d^r$ (Get acquainted with the notation)

where d = x – 4

Now, $\mu_1' = \frac{1}{N}\sum fd = 0$; $\quad \mu_2' = \frac{1}{N}\sum fd^2 = \frac{512}{256} = 2$

$\mu_3' = \frac{1}{N}\sum fd^3 = 0$; $\quad \mu_4' = \frac{1}{N}\sum fd^4 = \frac{2816}{256} = 11$

Moments about mean are $\mu_1 = 0$

$$\mu_2 = \mu_2' - \mu_1'^2 = 2 - 0 = 2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 0 - 0 + 0 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= 11 - 0 + 0 - 0 = 11$$

Now, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{4} = 0$ $\quad$ and $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2.75$

2.10 WIGHTED MEAN

If some items in the distribution are more important than others, then this point must be taken into consideration, while computing the mean of the distribution.

In such cases, the proper weightage must be given to various items, the weights attached to each item being proportional to the importance of the item in the distribution.

For example, if we want to have an idea of the change in cost of living of a certain group of people, then all the commodities are not equally important, e.g., wheat, rice and pulses are more important, e.g., wheat, rice and pulses are more important than cigarettes, tea, confectionaries etc.

Let w_i be the weight attached to the item xi, i – 1,2 , ...n. then we define :

Weignted A. M. (or weighted mean)

$$= \frac{\sum_i w_i x_i}{\sum_i w_i} \qquad ...(i)$$

We note that the formula for weighted mean is same as the formula for simple mean with f_i (i = 1, 2, ...n), the **frequencies replaced by** w_i, (i = 1, 2, ..., n), the weights.

Remarks :

(i) Weighted mean gives the result equal to the simple mean if the weights assigned to each of the variate values are equal.

(ii) It results in higher value than the simple mean if smaller weights are given to smaller items and larger weights to larger items.

(iii) If the weights attached to larger items are smaller and those attached to smaller items are larger, then the weighted mean results in smaller value than the sample mean.

Example : Find the simple and weighted arithmetic mean of the first n natural numbers, the weights being the corresponding numbers.

☑ **Soln. :**

The first n natural numbers are : 1, 2, 3, ...n

We have that

$1 + 2 + 3 + ... + n = \frac{n(n+1)}{2}$ and $1^2 + 2^2 + ...+ n^2 = \frac{n(n+1)(2n+1)}{6}$

X	W	W_X
1	1	1^2
2	2	2^2
3	3	3^2
⋮	⋮	⋮
N	N	N^2

$\therefore$ **Simple A. M.** $\bar{x} = \frac{\sum x}{n} = \frac{1 + 2 + ...+ n}{n} = \frac{n(n+1)}{2n}$

$= \frac{n+1}{2}$

and **weighted A. M.**

$$\bar{X}_w = \frac{\sum wX}{\sum w} = \frac{1^2 + 2^2 + ...+ n^2}{1 + 2 + ...+ n} = \frac{n(n+1)(2n+1)}{6} \cdot \frac{2}{n(n+1)} = \frac{2n+1}{3}$$

2.11 PROPERTIES OF ARTHMETIC MEAN

(1) **Property (1) :** Algebraic sum of the deviations of a set of values from their arithmetic mean is zero.

If (x_i, f_i), i = 1, 2, ...n is the frequency distribution, then

$\sum_{i=1}^{n} f_i (x_i - \bar{x}) = 0$, $\bar{x}$ is the mean of the distribution

Proof : we have

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$\therefore \sum_{i=1}^{n} f_i x_i = \bar{x} \sum_{i=1}^{n} f_i \quad ...(i)$$

Now, $\sum_{i=1}^{n} f_i (x_i - \bar{x}) = \sum_{i=1}^{n} f_i x_i - \bar{x} \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_i x_i - \sum_{i=1}^{n} f_i x_i = 0$ from (i)

(2) **Property (2) :** The sum of squares of the deviations of a set of values is minimum when taken about mean.

Proof :

▶ **Step (1) :** Let (x_i, f_i), i = 1, 2, ...n, be the frequency distribution.

$$\text{Let } Z = \sum_{i=1}^{n} f_i (x_i - A)^2,$$

be the sum of squares of deviations of given values from any arbitrary point A.

▶ **Step (II) :** We prove that Z is minimum for $A = \bar{x}$. we apply principle of maxima and minima.

(i) $Z = \sum_{i=1}^{n} f_i (x_i - A)^2$

Differentiating partially w.r.t. A,

$$\frac{\partial Z}{\partial A} = -2 \sum_{i=1}^{n} f_i (x_i - A) \qquad \text{Let } \frac{\partial Z}{\partial A} = 0 \Rightarrow \sum_{i=1}^{n} f_i (x_i - A) = 0$$

$$\therefore \sum_{i=1}^{n} f_i x_i - A \sum_{i=1}^{n} f_i = 0 \qquad \therefore A = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \bar{x}$$

$$\therefore A = \bar{x} \quad ...(ii)$$

▶ **Step (III) :** Differentiating $\frac{\partial Z}{\partial A}$ partially w.r.t. A,

$$\frac{\partial^2 Z}{\partial A^2} = -2(-1) = 2 > 0 \quad \therefore \frac{\partial^2 Z}{\partial A^2} > 0,$$

Hence Z is minimum at $A = \bar{x}$ hence, the result.

Property (3) : (mean of the composite series)

If $\bar{x}_i$ (i = 1, 2 ...., k) are the means of k component series of sizes n_i, (i = 1, 2, ..., k) respectively, then the mean $\bar{x}$ of the composite series obtained on combining the component series is given by :

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + ... + n_k \bar{x}_k}{n_1 + n_2 + ...+ n_k} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{n} n_i}$$

Proof

▶ **Step (I)**

Let $x_{11}, x_{12}, ...x_1 n_1$ be n_1 members of the first series; $x_{21}, x_{22}, ...x_2 n_2$ be the n_2 members of the second series, ..., $x_{k1}, x_{k2}, ..., x_k n_k$ be n_k members of the k^{th} series, then we have,

$$\bar{x}_1 = \frac{x_{11} + x_{12} + ... + x_1 n_1}{n_1}$$

$$\bar{x}_2 = \frac{x_{21} + x_{22} + ... + x_2 n_2}{n_2}$$

$$\vdots$$

$$\bar{x}_k = \frac{x_{k1} + x_{k2} + ... x_{knk}}{n_k} \quad ...(iii)$$

▶ **Step (II) :** The composite series is of size is of size

$(n_1 + n_2 + ... n_k)$,

The mean $\bar{x}$ of the composite series is,

$$\bar{x} = \frac{(x_{11} + x_{12} + ...+ x_1 n_1) + (x_{21} + x_{22} + ... + x_2 n_2) + ... + (x_{k1} + x_{k2} + ... + x_{knk})}{n_1 + n_2 + ... + n_k}$$

$$= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + ... + n_k \bar{x}_k}{n_1 + n_2 + ... + n_k} \quad ...\text{from (i)} \quad = \frac{\sum_i n_i x_i}{\sum_i n_i}$$

Ex. 2.11.1 : The average salary of male employees in a firm was Rs. 5200 and that of females was Rs. 4,200. The mean salary of all the employees was Rs. 5,000. Find the percentage of male and female employee.

☑ **Soln. :**

▶ **Step (1) :** Let n_1 and n_2 be respectively the number of male and female employees in the firm, and let $\bar{x}_1$ and $\bar{x}_2$ be respectively their average salary. Let $\bar{x}$ be the average salary of all the workers in the firm.

We have, $\bar{x}_1 = 5200, \bar{x}_2 = 4, 200, \bar{x} = 5000.$

▶ **Step (2) :** since $\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$

$$\therefore \quad 5000 = \frac{n_1 (5200) + n_2 (4200)}{n_1 + n_2}$$

$$\therefore \quad 50 (n_1 + n_2) = 52 n_1 + 42 n_2 \qquad \therefore 8 n_2 = 2 n_1$$

$$\therefore \frac{n_1}{n_2} = \frac{4}{1} \quad \text{i.e. } n_1 = 4 n_2$$

∴ The percentage of male employees in the firm

$$= \frac{4}{4+1} \times 100 = 80$$

and the percentage of female employees in the firm $= \frac{1}{(4+1)} \times 100 = 20$

2.12 MERITS AND DE-MERITS OF ARITHMETIC MEAN, MEDIAN, MODE, GEOMETRIC MEAN AND HARMONIC MEAN

(I) Merits and Demerits of Arithmetic Mean

Sr. No.	Merits	Demerits
(1)	If is rigidly defined.	It cannot be determined by mere inspection or by graphically.
(2)	Calculation is not complicated.	It fails to deals with qualitative characteristics e.g. intelligence, beauty, honesty etc.
(3)	It is based on mere observations.	It cannot be obtained if a single observation is missing or lost.
(4)	The mean of the composite series in terms of means and sizes of the component series is given by $\bar{x} = \frac{\sum_{i=1}^{n} n_1 \bar{x}_1}{\sum_{i=1}^{n} n_1}$	Extreme values affect the A.M. If the extreme values are high or very low, clear picture of the distribution is not possible to obtain.
(5)	Crosland to other averages, arithmetic mean is least affected by fluctuations of sampling. Hence A.M. is also termed as stable average.	Arithmetic means cannot be calculated if the extreme class is not mentioned, i.e. below 10 or above 90.

(II) Merits and Demerits of Median

Sr. No.	Merits	Demerits
(1)	It is rigidly defined.	If there are even number of observations, then median is the average of the two middle terms. Hence it cannot be determined exactly.
(2)	In many cases, it can be found by mere inspection.	Median is described as insensitive. For example median of 10, 20, 30, 40, 50 is 30. Put it we replace first two values 10, 20, by smaller then 30 and last two values 40, 50 by greater than 30, still median remains the same.
(3)	It is not affected by extreme values.	Algebraic treatment does not work.
(4)	It can be calculated for distribution with open-end classes.	As compared with mean, it is affected by fluctuations of sampling.

(III) Merits and Demerits of Mode

Sr. No.	Merits	Demerits
(1)	Mode is easy to calculate and is easily located.	It is not always possible to find a clearly defined mode. In some cases, there may be two modes in a given distributions. In that case, it is called as bi-modal. It can have more than two modes and then it is called as multi-modal.
(2)	Mode is not affected by extreme values.	Mode is not based upon all the observations.
(3)	Mode can be conveniently found even if the frequency distribution has class-intervals of unequal magnitude. Mode can be located even if there are open end classes.	Further mathematical treatment is not workable. Mode is affected by fluctuations of sampling.

(IV) Merits and Demerits of Geometric Mean

Sr. No.	Merits	Demerits
(1)	It is rigidly defined.	A person with no mathematical knowledge cannot so easily find G.M.
(2)	It is based upon all observations.	If any term is zero, then G.M. is 0 and if any term is negative, then G.M. becomes a complex number.
(3)	It is not much affected by fluctuations of sampling.	
(4)	In its treatment, small items are given more weightage.	

(V) Merits and Demerits of Harmonic Mean

Sr. No.	Merits	Demerits
(1)	Harmonic mean is rigidly defined.	Harmonic mean is not so easy to understand .
(2)	It is suitable for further mathematical treatment.	It is difficult numerically to compute.
(3)	It is not much affected by fluctuations of sampling.	
(4)	It is useful when small items have to be given a greater weightage.	

Chapter Ends...

❑❑❑