CHAPTER 1

CONCEPT OF ECONOMICS AND SIGNIFICANCE OF STATISTICS IN ECONOMICS

1. CONCEPT OF ECONOMICS

We advise the young learners of Class XI to comprehend the concept of economics through the following discussion relating to ordinary (or routine) business (or activities) of our life.

Every individual, ranging from a child to an old man, is engaged in some economic activity or the other. Consumption is an important economic activity; and we all are consumers, consuming different goods and services for the satisfaction of our wants.

Who is a Consumer?

A consumer is one who consumes goods and services for the satisfaction of his wants.

What is Consumption?

Consumption is the process of using up utility value of goods and services for the direct satisfaction of our wants. Utility value of goods means inherent capacity of goods and services to satisfy human wants.

Production is another economic activity, and many of us are producers, engaged in the production of different goods and services for the generation of income.

Who is a Producer?

A producer is one who produces and/or sells goods and services for the generation of income.

What is Production?

Production is the process of converting raw material into useful things. Things become useful as they acquire utility value in the process of production.

Saving and investment are also economic activities, and many of us are savers and investors. We save a part of our income for future consumption or for investment in shares and bonds to generate income.

What is Saving?

It is that part of income which is not consumed. It is an act of abstinence from consumption.

What is Investment?

It is expenditure by the producers on the purchase of such assets which help to generate income.

Thus, we are consumers, producers, savers and investors. We are all engaged in diverse economic activities. Economic activities include consumption, production, saving, investment, and many more.

What is Economic Activity?

It is an activity which is related to the use of scarce means (also called scarce resources). Means are always scarce in relation to our wants. Imagine yourself as the richest person on the earth. Still you can't have everything you wish to have at a point of time. It implies the scarcity of your means/ resources in relation to your wants.

Engaged in diverse economic activities, we are performing ‘ordinary business of life’, according to **Alfred Marshall,** a great pro founder of Modern Economics. Thus, he defines economics as “the study of mankind in the ordinary business of life."

Scarcity is the Undercurrent of Economic Problem and therefore of Economics

Resources are always scarce in relation to our wants. Also, resources have alternative uses: A ten-rupee note in your pocket may be spent on a cup of coffee or a cold drink. Likewise, a worker may render his services in factory A, rather than B and C. Because, resources are scarce and have alternative uses, we cannot escape from the problem of allocation of limited means to alternative uses. This is what we call economic problem or the problem of choice.

What is Economic Problem?

It is the problem of choice (or the problem of allocating scarce resources to alternative uses) arising on account of the fact that resources are scarce and these have alternative uses.

Economics is essentially the study of economic problems that we must confront owing to the fact that our means are scarce in relation to our wants, and that the means have alternative uses. If there is no scarcity, there is no economic problem, and there is no economics if there is no economic problem. Thus, **Robbins** defines economics as “A science that studies human behaviour as a relationship between ends and scarce means which have alternative uses.”

Three Distinct Components of Economics: Consumption, Production and Distribution

Consumption

Here, we, as students of economics, study behaviour of human beings as consumers or buyers of different goods and services for the satisfaction of their wants. As consumers, people have limited means, while their wants are unlimited. How do they allocate their given means (or income) to the purchase of different goods and services, (given their market prices) so that their satisfaction is maximised? This is the study of consumption or the study of consumer behaviour. When we formulate a set of standard relationships (like the inverse relationship between price of good and its purchase) explaining how consumers tend to behave, we call it consumption theory.

Production

Producers also have limited means while they have a wide range of goods and services to choose from for their firms and factories. Given prices of different inputs, how do they choose such combination(s) which are least expensive, so that they are able to minimise their cost of production. Also, given prices of different goods, how do they choose to produce those, the production of which offers them maximum revenue, so that their profit (profit = revenue - costs) is maximised. This is the study of production, or the study of producers’ behaviour. When we formulate a set of standard relationships (like greater the productivity of a factor, greater is its employment) explaining the behaviour of producers or their production decisions, we call it production theory.

Distribution

As students of economies we are also interested in knowing how income (generated in the process of production) is distributed among those who have worked as agents of production. Who are agents of production? These are owners of factors of production, viz. land, labour, capital and entrepreneurship. A part of income generated will go to the owners of land (used in production) in the form of rent; a part will go to labourers (for rendering their services) in the form of wage; a part will go to the owners of capital (used in production) in the form of interest; and a part will go to the entrepreneurs in the form of profits. Distribution of income refers to the distribution of GDP (gross domestic product) among the owners of the factors of production (land, labour, capital and entrepreneurship). What are the economic principles on the basis of which income is distributed among owners of the factors of production? Such a study is called distribution theory in economics.

Besides these three major components of economics, the economists also address such questions which are of social significance, like the question of poverty and unemployment, the question of growth with social justice and the question of environmental degradation as linked to various economic activities. Issues of social significance or collective significance are categorised as issues of macroeconomics. These are distinct from the issues of microeconomics which revolve around the problems of choice confronted by microeconomic units like a household, a firm or an industry.

Microeconomics and Macroeconomics

Microeconomics deals with economic issues or economic problems related to microeconomic units like a household, a firm or an industry. These issues and problems are studied and addressed largely with a view to maximising individual welfare. Macroeconomics deals with economic issues or economic problems at the level of economy as a whole. These issues or problems are studied and addressed keeping in mind the goals of social welfare or collective welfare.

2. WHAT IS STATISTICS?

Even to a layman this should not be a difficult question. If asked to define Statistics, we can expect a layman to say that Statistics is something like a store of quantitative information. **Yes, it is true.** Statistics means quantitative information or quantification of the facts and findings. But, how do we get quantitative information? There must be a system, method or technique to collect quantitative information. Also, statistical information may be a raw information. It needs to be classified, tabulated or it needs to be systematically presented. One must learn the system of presentation and classification of data. Also, there must be a set of methods and techniques to condense the data. May be, we find averages or percentages. And above all, there must be a set of methods or techniques on the analysis and interpretation of quantitative information. A student of economics has to study all these methods and techniques to understand and master the subject matter of Statistics.

Thus, unlike a layman, a student of economics cannot relax taking Statistics just as a pool of quantitative information. Instead he is also to look into the methods or techniques relating to its collection, classification, presentation, analysis as well as interpretation.

In view of such a vastness of the subject matter, Statistics is defined both in **singular sense** and **plural sense,** as under:

Statistics—A Plural Noun

In its plural sense, Statistics refers to information in terms of numbers or numerical data, such as Population Statistics, Employment Statistics, Statistics concerning Public Expenditure, etc. However, any numerical information is not Statistics. **Example:** Ram gets Rs. 100 per month as pocket allowance is not Statistics (it is neither an aggregate nor an average) whereas average pocket allowance of the students of Class X is Rs. 100 per month, or there are 80 students in Class XI compared to just 8 in Class XII of your school are Statistics.

The following table shows a set of data which is Statistics, and another set which is not Statistics. The figures used are hypothetical.

Data which are not Statistics	Data which are Statistics
(i) A cow has 4 legs.	(i) Average height of the 26-plus male people in India is 6 feet compared to 5 feet in Nepal.
(ii) Ram has 200 rupees in his pocket.	(ii) Birth rate in India is 18 per thousand compared to 8 per thousand in USA.
(iii) A young lady was run over by a speeding truck at 100 km per hour.	(iii) Over the past 10 years, India has won 60 test matches in cricket and lost 50.

Distinction between Quantitative and Qualitative Data

This is related to the distinction between quantitative variables and qualitative attributes. There are quantitative variables like income, expenditure and investment which can be expressed in numerical terms, viz., per capita income in India was (say) Rs. 40,000 per month, per capita expenditure was (say) Rs. 30,000 per month, and net investment (or capita formation) was (say) Rs. 10,000 crore in the year 2017-18. All such data are called quantitative data. On the other hand, there are qualitative attributes like 'IQ' level of different individuals or beauty of the individuals which cannot be expressed in numerical terms.

These attributes refer to qualitative characteristics of the objects. These can be ranked or rated as good, very good, or excellent. We may give them ranks as 1, 2, 3, etc. All such data are called qualitative data. Briefly, while quantitative data refers to quantitative variables, qualitative data refers to qualitative attributes of the different objects.

Thus:

All Statistics are data, but all data are not Statistics.

Definition

In its plural sense, this is how Statistics is defined by different authors:

“Statistics are numerical statements of facts in any department of enquiry placed in relation to each other. ” **—Bowley**

“By Statistics we mean quantitative data affected to a marked extent by multiplicity of causes. ” **—Yule and Kendall**

Features or Characteristics of Statistics in the Plural Sense or as Numerical Data

Main characteristics of Statistics in terms of numerical data are as follows:

(1) Aggregate of Facts: A single number does not constitute Statistics. No conclusion can be drawn from it. It is only the aggregate number of facts that is called Statistics, as the same can be compared and conclusions can be drawn from them. For example, if it is stated that there are 1,000 students in our college, then it has no statistical significance. But if it is stated that there are 300 students in arts faculty, 400 in commerce faculty and 300 in science faculty in our college, it makes statistical sense as this data conveys statistical information. Similarly, if it is stated that population of India is 121 crore or that the value of total exports from India is Rs. 14,41,420 crore, then these aggregate of facts will be termed as Statistics. It can, therefore, be concluded ‘All Statistics are expressed in numbers but all numbers are not Statistics’.

(2) Numerically Expressed: Statistics are expressed in terms of numbers. Qualitative aspects like ‘small’ or 'big’; ‘rich’ or ‘poor’; etc. are not called Statistics. For instance, to say, Irfan Pathan is tall and Sachin is short, has no statistical sense. However, if it is stated that height of Irfan Pathan is 6 ft and 2-inches and that of Sachin is 5 ft and 4-inches, then these numericals will he called Statistics.

(3) Multiplicity of Causes: Statistics are not affected by any single factor; but are influenced by many factors. Had they been affected by one factor alone then by removing that factor they would lose all their significance. For instance, 30 per cent rise in prices may have been due to several causes, like reduction in supply, increase in demand, shortage of power, rise in wages, rise in taxes, etc.

(4) Reasonable Accuracy: A reasonable degree of accuracy must be kept in view while collecting statistical data. This accuracy depends on the purpose of investigation, its nature, size and available resources.

(5) Mutually related and Comparable: Such numericals alone will be called Statistics as are mutually related and comparable. Unless they have the quality of comparison they cannot be called Statistics.

For example, if it is stated "Ram is 40 years old, Mohan is 5 ft tall, Sohan has 60 kg of weight", then these numbers will not be called Statistics, as they are not mutually related nor subject to comparison. However, if the age, height and weight of all the three are inter-related, then the same will be considered as Statistics.

(6) Pre-determined Objective: Statistics are collected with some pre-determined objective. Any information collected without any definite objective will only be a numerical value and not Statistics. If data pertaining to the farmers of a village is collected, there must be some pre-determined objective. Whether the Statistics are collected for the purpose of knowing their economic position, or distribution of land among them or their total population, or for any other purpose, all these objectives must be pre-determined.

(7) Enumerated or Estimated: Statistics may be collected by enumeration or are estimated. If the field of investigation is vast, the procedure of estimation may be helpful. For example, 1 lakh people attended the rally addressed by the Prime Minister in Delhi and 2 lakh in Mumbai. These Statistics are based on estimation. As against it, if the field of enquiry is limited, the enumeration method is appropriate. For example, it can be verified by enumeration whether 20 students are present in the class or 10 workers are working in the factory.

(8) Collected in a Systematic Manner: Statistics should be collected in a systematic manner. Before collecting them, a plan must be prepared. No conclusion can be drawn from Statistics collected in haphazard manner. For instance, data regarding the marks secured by the students of a college without any reference to the class, subject, examination or maximum marks, etc., will lead to no conclusion.

In short, it can safely be concluded that "all numerical data cannot be called Statistics but all Statistics are called numerical data. "

Statistics-A Singular Noun

In the singular sense, Statistics means science of Statistics or statistical methods. It refers to techniques or methods relating to collection, classification, presentation, analysis and interpretation of quantitative data.

Focus of the Study

Statistics as a singular noun is focus of the study for the students of Class XI. You are to learn and understand how to collect data, organise data, present data as well as analyse and interpret data.

Definition

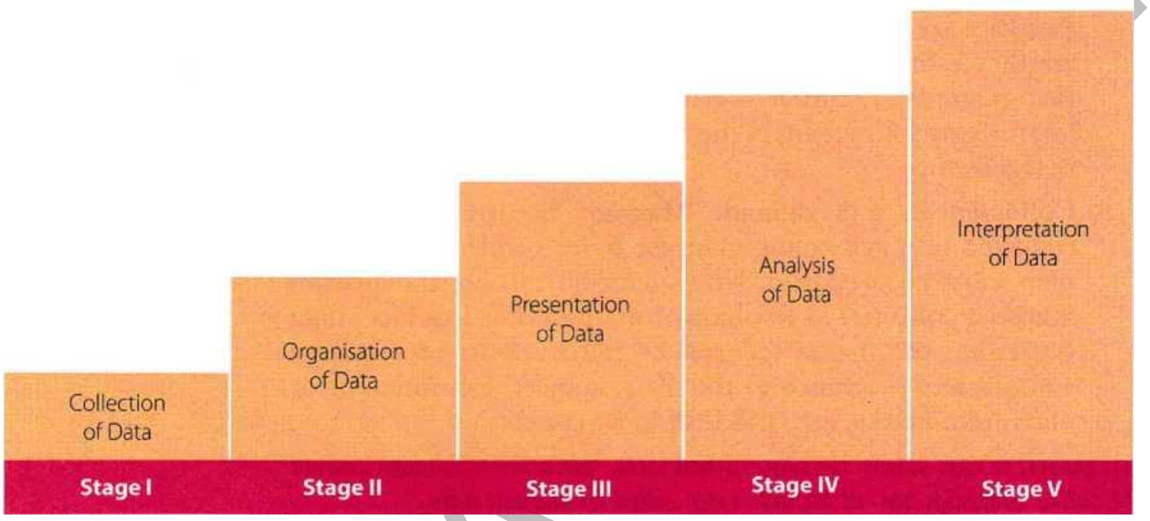
"Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data. " **—Croxton and Cowden**

“Statistics is the science which deals with the collection, classification and tabulation of numerical facts as a basis for the explanation, description and comparison of phenomena. ” **—Lovitt**

“Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and. interpreting numerical data, collected to th row some light on any sphere of enquiry. ” **—Seligman**

Stages of Statistical Study

Studying Statistics as a singular noun implies the knowledge of various stages of statistical study. These stages are:



Obviously, at the first stage, we collect statistical data. Second, we organise the data in some systematic order. Third, we present the data in the form of graphs, diagrams or tables. Fourth, we analyse the data in terms of averages or percentages. Fifth, and finally, we interpret the data to find certain conclusions.

Statistical Tools

Each stage of the statistical study involves the use of certain standard techniques or methods. These techniques or methods are called statistical tools. Thus, there are statistical tools used for the collection of data, like the ‘Sample’ and ‘Census’ techniques. Array of data and tally bars are the standard techniques used for organisation of data. Tables, graphs and diagrams are the well-known statistical tools for the presentation of data.

Averages and percentages are the commonly used techniques for the analysis of data. Interpretation of data is often done in terms of the magnitude of averages, percentages or coefficients of correlation/regression. The following table gives an overall view of the various stages of statistical study and the related sets of statistical tools.

What are Statistical Tools?

These refer to the methods or techniques used for the collection, organisation and presentation of data, as well as for the analysis and interpretation of data.

Stages of Statistical Study and the Related Statistical Tools

Stages	Statistical Study	Statistical Tools
Stage I	Collection of Data	Census or Sample Techniques
Stage II	Organisation of Data	Array of Data and Tally Bars
Stage III	Presentation of Data	Tables, Graphs and Diagrams
Stage IV	Analysis of Data	Percentages, Averages, Correlation and Regression Coefficients
Stage V	Interpretation of Data	Magnitude of Percentages, Averages and the Degree of Relationship between different economic variables

3. SCOPE OF STATISTICS

Study of the scope of statistics includes:

(1) Nature of Statistics,

(2) Subject Matter of Statistics, and

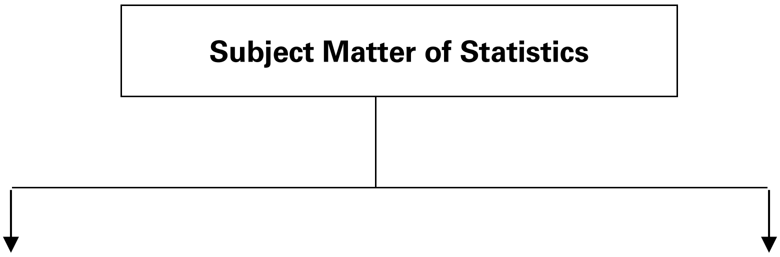
(3) Limitations of Statistics.

Nature of Statistics

Here, the basic question is whether Statistics is a science or an art. **Prof. Tippet,** has rightly observed that "Statistics is both a science as well as an art." As a science, Statistics studies numerical data in a scientific or systematic manner. As an art, Statistics relates to quantitative data to the real life problems. By using statistical data, we are able to analyse and understand real life problems much better than otherwise. Thus, the problem of unemployment in India is more meaningfully analysed when the size of unemployment is supported with quantitative data.

Subject Matter of Statistics

Subject matter of statistics includes two components: Descriptive Statistics and Inferential Statistics.



Subject Matter of Statistics

(1) Descriptive Statistics	(2) Inferential Statistics

The Concept of Universe or Population

It should be interesting for the students of Clsss XI to note that the concept of universe or population has a specific meaning in Statistics. It refers to the aggregate of all items or units relating to your statistical study. Example:

Universe or population size is 1,000 if you are studying 1,000 students for your statistical study.

(1) **Descriptive Statistics:** Descriptive Statistics refers to those methods which are used for the collection, presentation as well as analysis of data. These methods relate to such estimations as ‘measurement of central tendencies’ (average mean, median, mode), ‘measurement of dispersion’ (mean deviation, standard deviation, etc.), 'measurement of correlation’, etc. **Example:** Descriptive statistics is used when you estimate average height of the secondary students in your school. Likewise, descriptive statistics is used when you find that marks in science and mathematics of the students in all classes are intimately related to each other.

(2) Inferential Statistics: Inferential Statistics refers to all such methods by which conclusions are drawn relating to the universe or population on the basis of a given sample. (In Statistics, the term universe or population refers to the aggregate of all items or units relating to any subject.) For example, if your class teacher estimates average weight of the entire class (called universe or population) on the basis of average weight of only a sample of students of the class, he is using inferential statistics.

Limitations of Statistics

In modern times. Statistics has emerged to be of crucial significance in all walks of life. However, it has certain limitations. Thus, writes **Newshome** that, “Statistics must be regarded as an instrument of research of great value but barring severe limitations which are not possible to overcome." Following are some notable limitations of Statistics:

(1) Study of Numerical Facts only: Statistics studies only such facts as can be expressed in numerical terms, it does not study qualitative phenomena like honesty, friendship, wisdom, health, patriotism, justice, etc.

(2) Study of Aggregates only: Statistics studies only the aggregates of quantitative facts. It does not study statistical facts relating to any particular unit. **Example:** It may be a statistical fact that your class teacher earns Rs. 50,000 per month. But, as this fact relates to an individual, it is not to be deemed as a subject matter of Statistics. However, it becomes a subject matter of Statistics if we study income of school teachers across all parts of the country, for purpose of finding regional differences in income.

(3) Homogeneity of Data, an essential Requirement: To compare data, it is essential that statistics are uniform in quality. Data of diverse qualities and kinds cannot be compared. For example, production of food grains cannot be compared with the production of cloth. It is because cloth is measured in meters and food grains in tonnes. Nevertheless, it is possible to compare their value instead of the volume.

(4) Results are True only on an Average: Most statistical findings are true only as averages. They express only the broad tendencies. Unlike the laws of natural sciences, statistical observations are not error-free. They are not always valid under all conditions. For instance, if it is said that per capita income in India is Rs. 50,000 per annum, it does not mean that the income of each and every Indian is Rs. 50,000 per annum. Some may have more and some may have less,

(5) Without Reference, Results may Prove to be Wrong: In order to understand the conclusions precisely, it is necessary that the circumstances and conditions under which these conclusions have been drawn are also studied. Otherwise, they may prove to be wrong.

(6) Can be used only by the Experts: Statistics can be used only by those persons who have special knowledge of statistical methods. Those who are ignorant about these methods cannot make sensible use of statistics. It can, therefore, be said that data in the hands of an unqualified person is like a medicine in the hands of a quack who may abuse it, leading to disastrous consequences. In the words of **Yule** and **Kendall, “**Statistical methods are most dangerous tools in the hands of an inexpert.”

(7) Prone to Misuse: Misuse of Statistics is very common. Statistics may used to support a pre-drawn conclusion even when it is absolutely false. It is usually said, “Statistics are like clay by which you can make a god or a devil, as you please.” Misuse of statistics is indeed its greatest limitation.

4. FUNCTIONS AND IMPORTANCE OF STATISTICS IN ECONOMICS

Following words of **Prof. Tippet** very aptly capture the importance of Statistics in economics: “A day might come when the department of economics in the universities will go out of the control of economic theoreticians and come under the control of statistical workshops, in the same manner as the department of physics and chemistry have come under the control of experimental laboratories.” Indeed, Statistics has emerged as the lifeline of economics. It is because of the growing use of Statistics by the economists that the subjects like econometrics have been added to the horizons of economics. Students of Class XI may note the following points to highlight the significance (functions and importance) of Statistics in economics.

(1) Quantitative Expression of Economic Problems:

Consider any economic problem, be it the problem of unemployment, the problem of price rise or the problem of shrinking exports. The first task of the economists is to understand its magnitude through its quantitative expression.

For example, if it is the problem of unemployment, we make its quantitative expression stating that (say) 20 per cent of the India’s working population is unemployed or that between the years 1995-2010 the percentage of unemployed working population has tended to increase from 18 per cent to 9.4 per cent.

(2) Inter-sectoral and Inter-temporal Comparisons:

Economists do not stop merely at the quantitative expression of the problems. They would try to further comprehend it through inter-sectoral and inter-temporal comparisons. From inter-sectoral comparisons we mean, comparisons across different sectors of the economy. Thus, analysing the problem of unemployment, the economists would like to know the magnitude of unemployment across rural and urban sectors of the economy. They would like to know what percentage of rural population is unemployed compared to the urban population. Inter-temporal comparison means understanding of change in the magnitude of the problem over time. This would mean making a comparison (say) over different plan periods of the rural and urban unemployment.

(3) Working out Cause and Effect Relationship:

Economists try to find out cause and effect relationship between different sets of data. This enables them to attempt an effective diagnosis of the problem and accordingly to suggest some effective remedies. Thus, through their statistical studies, if the economists come to know that it is because of the decline in demand that investment in the economy has tended to shrink, they can suggest the government to adopt such measures as would increase the level of demand in the economy.

Two Important points on the Significance of Statistics in Economics

(i) Statistics facilitates inter-sectoral and inter-temporal comparison.

(ii) Statistics helps to establish cause and effect relationship between different economic variables that have facilitated the construction of economic theories.

(4) Construction of Economic Theories or Economic Models:

What is economic theory? It is an established statistical relationship between different sets of statistical data, offering conclusions of economic significance. The well-known inverse relationship between price of a commodity and its demand (i.e., more is purchased when price falls) is an established statistical relationship, and therefore, is a part of economic theory. Is the construction of theoretical relationships or models possible without statistical experiments? Certainly not.

(5) Economic Forecasting: Economists do forecasting through statistical studies. By the term forecasting we do not mean some astrological predictions. We only mean to assess and ascertain the future course of certain events which are of economic significance. Thus, on studying the behaviour of price level over several years, the economists can make statistical forecasting about the likely trend or pattern of the price level in the near future. This helps us in future planning.

(6) Formulation of Policies: How does the finance minister decide to increase or decrease taxation as a source of government revenue? Obviously through statistical studies. It is through statistical investigations that the finance minister gets a feedback on the tax-paying capacity of the people, and revenue needs of the government. Accordingly, tax rates are fixed to get maximum possible revenue with minimum possible discomfort to the people.

(7) Economic Equilibrium: What is economic equilibrium? It is a state of balance for the producer or the consumer where the producer finds that his profits are maximum or where the consumer finds that his satisfaction is maximum. It is through the use of statistical methods that the economists have evolved some eco-fundamentals (which you will study in Class XII) telling us how profits of the producers are maximised or how consumers get maximum satisfaction.

Thus, so much is the significance of Statistics in economics that **Marshall** (a great economist of the past century) had to concede that “Statistics are the straw out of which I, like every other economist, have to make bricks." Surely, Statistics is the hub of the wheel of economic studies, and the beginners of Class XI must focus on the hub to precisely understand the movement of the entire wheel.

Statistical Methods are No Substitute for Common Sense

This is a statement of caution to the students of Statistics. It urges the students not to use Statistics devoid of their common sense. You may find some spurious relationships, like larger the number of doctors in an area greater are the deaths in that area. It may be true statistically, but does not match with common sense. Hence, never propagate any statistical conclusion in case it offends your common sense. Likewise, average size of shoes for the 50 students in your class may be 'six'. But it would be foolish if the school authorities place an order of 50 shoes of the size-six for all of you. Surely this size may not fit many of you.

Distrust of Statistics

Some people have misgivings about Statistics and make observations like the following:

(i) Statistics is a rainbow of lies.

(ii) Statistics are tissues of falsehood.

(iii) Statistics can prove anything.

(iv) Statistics cannot prove anything.

(v) Statistics are like clay of which you can make a god or a devil, as you please.

According to Disraeli, "There are three kinds of iies-lies, damned lies and Statistics."

Indeed, one can present statistical information in a manner that tends to distort the facts and thereby mislead the people. For instance, the government claimed that in 2018, per capita income in India increased by about 17 per cent. On the other hand, the opposition party claimed that in 2018, per capita income increased by 5 per cent only. But the difference lies in the fact that whereas government estimates are based on current prices, those of the opposition party are based on the 2011-12 prices, it is difficult for a layman to understand this difference. He will just be confused or perhaps be fooled by the claims and counterclaims of the government and the opposition party.

What Causes Distrust?

Distrust of Statistics arises not because there is anything wrong with Statistics as a subject matter. It arises because the users of Statistics tend to manipulate it to suit or support their pre-drawn conclusions or observations. Main causes for the distrust of Statistics are as under:

(i) Different kinds of Statistics are obtained in respect of a given problem.

(ii) Statistics can be altered to match the predetermined conclusions.

(iii) Authentic Statistics can also be presented in such a manner as to confuse the reader.

(iv) When Statistics are collected in a partial manner, the results are generally wrong. Consequently, people lose faith in them.

However, it may be noted that if Statistics are presented wrongly, then the fault does not lie with Statistics as a subject matter. The fault lies with those people who collect wrong Statistics or those who draw wrong conclusions. **Statistics, as such, do not prove anything. They are simply tools in the hands of the statisticians. If a statistician misuses the data, then the blame lies squarely on him and not on the subject matter, A competent doctor can cure a disease by making good use of the medicine but the same medicine in the hands of an incompetent doctor becomes a poison. The fault in this case is not of the medicine but of the unqualified doctor. In the same way, Statistics is never faulty but the fault lies with the users.**

In fact, Statistics should not be relied upon blindly nor distrusted outright. “Statistics should not be used as a blind man uses a lamp post for support rather than for illumination, whereas its real purpose is to serve as illumination and not as a support.”

In making use of Statistics one should be cautious and vigilant. In the words of King, "The science of Statistics is the most useful servant, but only of great value to those who understand its proper use.”

It is the duty of the students of economics to make use of know-how of Statistics to discover the truth rather than to cover the truth.

How to Remove Distrust?

Following are some essential remedies of the distrust of Statistics:

(i) Consideration of Statistical Limitations: While making use of Statistics, limitations of Statistics must be taken care of.

(ii) No Bias: The user should be impartial. He should make use only of the relevant data and draw conclusions without any bias or prejudice.

(iii) Application by Experts: Statistics should be used only by the experts to minimise the possibility of misuse.

Multiple Choice Questions

Select the correct alternative:

1. Which of the following statements is not an example of Statistics?

(a) Birth rate in India is 18 per thousand as compared to 8 per thousand in the US.

(b) Ramesh has a Rs. 100 note in his pocket.

(c) Over the last 10 years, India has won 60 text matches in cricket and lost 50.

(d) Average pocket allowance of the students of Class XI is Rs. 500 per month.

2. Which of the following is correct regarding Statistics?

(a) Aggregate of facts

(b) Numerically expressed

(c) Affected by multiplicity of causes

(d) All of these

3. In singular sense Statistics means:

(a) statistical science

(b) statistical law

(c) both (a) and (b)

(d) none of these

4. The aggregate of data is called:

(a) Statistics

(b) editing of data

(c) analysis of data

(d) collection of data

5. Which of the following indicates a stage of statistical study?

(a) Collection of data

(b) Presentation of data

(c) Analysis of data

(d) All of these

6. In plural sense, which of the following is not a characteristic of Statistics?

(a) Aggregate of data

(b) Only expressed in words

(c) Affected by multiplicity of causes

(d) Collected in a systematic manner

7. With regards to distrust of Statistics which of the following statements is not correct?

(a) Statistics is a rainbow of lies

(b) Statistics are tissues of falsehood

(c) Statistics express the facts in numbers

(d) There are three kinds of lies-lies, damned lies and statistics

8. Which of the following is an economic activity?

(a) Production

(b) Consumption

(c) Distribution

(d) All of these

9. Which of the following statements is incorrect?

(a) Resources have alternative uses

(b) All numbers are Statistics

(c) Macroeconomics studies large aggregates

(d) Statistics studies only the aggregates of quantitative facts

10. The process of converting raw material into goods is called:

(a) production

(b) saving

(c) investment

(d) exchange

11. The word 'statistics' is used as:

(a) Singular

(b) Plural

(c) Singular and Plural both

(d) None

12. The proper use of statistics can be made by:

(a) Cheats

(b) Everyone

(c) Experts

(d) Common Man

13. Statistics as a singular noun means:

(a) Statistical data

(b) Statistical methods

(c) Inductive statistics

(d) Descriptive statistics

14. Statistics is the science of analysing:

(a) Qualitative data

(b) Quantitative data

(c) Any kind of data

(d) Both (a) and (b)

15. Statistics as a plural noun indicates:

(a) Statistical methods

(b) Descriptive statistics

(c) Statistical data

(d) Inductive statistics

16. The statistics is concerned with:

(a) Aggregate of organised facts

(b) Aggregate of disorganised facts

(c) Aggregate of useless facts

(d) Aggregate of unrelated facts

17. Distrust of statistics is due to:

(a) Misuse of statistics

(b) Insufficient statistical methods

(c) Scope of statistics is limited

(d) Limitations of statistics

18. Statistics in singular sense includes:

(a) Collection of data

(b) Organisation of data

(c) Presentation of data

(d) All of the above

19. Statistics is defined in terms of numerical data in the:

(a) Singular Sense

(b) Plural Sense

(c) Either (a) or (b)

(d) Both (a) and (b)

20. The word ‘statistics’ is used as:

(a) Singular

(b) Plural

(c) Singular and Plural both

(d) None

21. The proper use of statistics can be made by:

(a) Cheats

(b) Everyone

(c) Experts

(d) Common Man

22. Statistics as a singular noun means:

(a) Statistical data

(b) Statistical methods

(c) Inductive statistics

(d) Descriptive statistics

23. Statistics is the science of analysing:

(a) Qualitative data

(b) Quantitative data

(c) Any kind of data

(d) Both (a) and (b)

24. Statistics as a plural noun indicates:

(a) Statistical methods

(b) Descriptive statistics

(c) Statistical data

(d) Inductive statistics

25. The statistics is concerned with:

(a) Aggregate of organised facts

(b) Aggregate of disorganised facts

(c) Aggregate of useless facts

(d) Aggregate of unrelated facts

26. Distrust of statistics is due to:

(a) Misuse of statistics

(b) Insufficient statistical methods

(c) Scope of statistics is limited

(d) Limitations of statistics

27. Statistics in singular sense includes:

(a) Collection of data

(b) Organisation of data

(c) Presentation of data

(d) All of the above

28. Statistics is defined in terms of numerical data in the:

(a) Singular Sense

(b) Plural Sense

(c) Either (a) or (b)

(d) Both (a) and (b)

Answers

Multiple Choice Questions

1. (b)	**2. (d)**	**3. (c)**	**4. (a)**	**5. (d)**
6. (b)	**7. (c)**	**8. (d)**	**9. (b)**	**10. (a)**
11. (a)	**12. (c)**	**13. (b)**	**14. (b)**	**15. (c)**
16. (a)	**17. (a)**	**18. (d)**	**19. (b)**	**20. (c)**
21. (c)	**22. (b)**	**23. (b)**	**24. (c)**	**25. (a)**
26. (a)	**27. (d)**	**28. (b)**		

CHAPTER 2

COLLECTION OF DATA

Statistics means data or quantitative information capable of some meaningful conclusions. The present chapter focuses on the collection of data, including:

(i) sources of data, and

(ii) methods of collecting data.

The purpose of data collection is to understand, analyse and explain a socio-economic problem, for example, the problem of unemployment or the problem of poverty. While analysing the problem we also try to understand the causes behind the problem as also the possible solutions. The entire exercise needs a comprehensive support of numerical facts, called data.

1. PRIMARY AND SECONDARY SOURCES OF COLLECTION OF DATA

There are two sources of collection of data:

(i) Primary Source, and

(ii) Secondary Source.

Primary Source:

You want to know about the quality of life of the people in your town. You may like to ascertain the quality of life in terms of per capita expenditure of different households in your town. You decide to collect the basic data yourself through statistical survey(s), of course with the help of investigators or field workers. While doing this exercise you are relying on primary source of the data. Thus, primary source of data implies collection of data from its source of origin. It offers you firsthand quantitative information relating to your statistical study. You or your team of investigators are contacting the respondents (people offering basic information) and obtaining the desired quantitative information on per capita expenditure of different households in your town.

Primary source of data implies collection of data from its source of origin. It offers you first-hand quantitative information relating to your statistical study.

Secondary Source:

Secondary Source of collection of data implies obtaining the relevant statistical information from an agency, or an institution which is already in possession of that information. To continue with the previous example, data relating to the quality of life of the people of your town (or the data on per capita expenditure) may have already been collected by the State Government. You can simply approach the concerned Government department and request for the desired information. This will be a Secondary Source of data for you. Thus, secondary source implies that the desired statistical information already exists and you are simply to collect it from the concerned agency or the department. You are not to conduct statistical survey(s) yourself and you are not to contact the respondents (people offering basic information). OT course, you are not getting first hand information relating to your statistical study. You are simply relying on the information which is already existing.

Secondary source of data implies collection of data from some agency or institution which already happens to have collected the data through statistical survey(s). It does not offer you first-hand information relating to your statistical study. You are to rely on the information which is already existing.

Primary and Secondary Data

Primary source of collecting data offers you, 'Primary Data' and secondary source offers you Secondary Data'. Let us clarify the difference.

Primary Data: Data collected by the investigator for his own purpose, for the first time, from beginning to end, are called primary data. These are collected from the source of origin. In the words of **Wessel,** "Data originally collected in the process oj investigation are known as primary data." Primary data are original. The concerned investigator is the first person who collects this information. The primary data are, therefore, a first-hand information. To illustrate, you may be interested in studying the socio-economic state of those students in your Class XI who secured first division in their matriculation examination. You collect information regarding their pocket allowance, their family income, educational status, their family members and the like. All this information would be termed as primary information or primary data, since you happen to be the first person to collect this information from the source of its origin.

Secondary Data: In the words of **M.M. Blair,** "Secondary data are those which are already in existence, and which have been collected, for some other purpose than the answering of the question in hand." According to **Wessel,** "Data collected by other persons are called secondary data." These data are, therefore, called second-hand data. Obviously, since these data have already been collected by somebody else, these are available in the form of published or unpublished reports. For example, data relating to Indian Railways which are annually published by the Railway Board, would be secondary data for any researcher.

Principal Differences between Primary and Secondary Data

The following are some principal differences between primary and secondary data:

(1) **Difference in Originality**: Primary data are original because these are collected by the investigator from the source of their origin. Against this, secondary data are already in existence and therefore, are not original.

(2) **Difference in Objective**: Primary data are always related to a specific objective of the investigator. These data, therefore, do not need any adjustment for the concerned study. On the other hand, secondary data have already been collected for some other

purpose. Therefore, these data need to be adjusted to suit the objective of study in hand.

(3) **Difference in Cost of Collection:** Primary data are costlier in terms of time, money and efforts involved than the secondary data. This is because primary data are collected for the first time from their source of origin. Secondary data are simply collected from the published or unpublished reports. Accordingly, these are much less expensive.

Of course, it may be noted that, there are no fundamental differences between primary data and secondary data. Data are data, whether primary or secondary. These are classified as primary or secondary just on the basis of their collection: first-hand or second-hand. Thus, a particular set of data when collected by the investigator for a specific purpose from the source of origin, would be primary data. And the same set of data, when used by some other investigator for his own purpose, would be known as secondary data. Thus, **Secrist** has rightly pointed out, "The distinction between primary and secondary data is one of the degree. Data which are primary in the hands of one party may be secondary in the hands of other.''

Primary and Secondary Data—The Basic Difference

■ If we are collecting data from its source of origin, for the first time, it is primary data.

■ If we are using data which have already been collected by somebody else, it is secondary data.

Note: If you are getting data from somebody else who collected it from its source of origin but did not use it for his own study, it will be deemed as primary data.

2. HOW BASIC DATA IS COLLECTED: SOME STATISTICAL METHODS/ MODES OF DATA COLLECTION

When basic data is to be collected from its primary source, how do we do it? It involves the study of a set of statistical methods or statistical techniques. The following are some of the well-known methods of collecting primary data:

(1) Direct Personal Investigation,

(2) Indirect Oral Investigation,

(3) Information from Local Sources or Correspondents,

(4) Information through Questionnaires and Schedules

(i) Mailing Method, and (ii) Enumerator's Method.

(1) Direct Personal Investigation

The direct personal investigation is the method by which data are personally collected by the investigator from the informants. In other words, the investigator establishes direct relation with the persons from whom the information is to be obtained. The success of this method, however, requires that the investigator should be very diligent, efficient, impartial and tolerant.

Direct contact with the workers of an industry to obtain information about their economic conditions is an example of this method.

Suitability

This method of collecting primary data is suitable particularly when:

(i) the field of investigation is limited or not very large.

(ii) a greater degree of originality of the data is required.

(iii) information is to be kept secret.

(iv) accuracy of data is of great significance, and

(v) when direct contact with the informants is required.

Merits

Data, thus, collected have the following merits:

(i) Originality: Data have a high degree of originality.

(ii) Accuracy: Data are fairly accurate when personally collected.

(iii) Reliability: Because the information is collected by the investigator himself, reliability of the data is not doubted.

(iv) Related Information: When in direct contact with the informants, the investigator may obtain other related information as well.

(v) Uniformity: There is a fair degree of uniformity in the data collected by the investigator himself from the informants. It facilitates comparison.

(vi) Elastic: This method is fairly elastic because the investigator can always make necessary adjustments in his set of questions.

Demerits

However, the method of direct personal investigation suffers from certain demerits, as under:

(i) Difficult to Cover Wide Areas: Direct personal investigation becomes very difficult when the area of the study is very wide.

(ii) Personal Bias: This method is highly prone to personal bias of the investigator. As a result, the data may lose their credibility.

(iii) Costly: This method is very expensive in terms of the time, money and efforts involved.

(iv) Limited Coverage: In this method, area of investigation is generally small. The results are, therefore, less representative. This may lead to wrong conclusions.

Learning by doing

You are to conduct direct personal investigation on the quality of teaching in a school. Who are your informants? What difficulties do you expect to encounter in such an exercise?

(2) Indirect Oral Investigation

Indirect oral investigation is the method by which information is obtained not from the persons regarding whom the information is needed. It is collected orally from other persons who are expected to possess the necessary information, these other persons are known as witnesses. For example, by this method, the data on the economic conditions of the workers may be collected from their employers rather than the workers themselves.

Suitability

This method is suitable particularly when:

(i) the field of investigation is relatively large.

(ii) it is not possible to have direct contact with the concerned informants.

(iii) the concerned informants are not capable of giving information because of their ignorance or illiteracy.

(iv) investigation is so complex in nature that only experts can give information.

This method is mosdy used by government or non-government committees or commissions.

Merits

Some of the notable merits of this method are as under:

(i) Wide Coverage: This method can be applied even when the field of investigation is very wide.

(ii) Less Expensive: This is relatively a less expensive method as compared to Direct Personal Investigation.

(iii) Expert Opinion: Using this method an investigator can seek opinion of the experts and thereby can make his information more reliable.

(iv) Free from Bias: This method is relatively free from the personal bias of the investigator.

(v) Simple: This is relatively a simple approach of data collection.

Demerits

However, there are some demerits, as under:

(i) Less Accurate: The data collected by this method are relatively less accurate. This is because the information is obtained from persons other than the concerned informants.

(ii) Biased: There is possibility of personal bias of the witnesses giving information.

(iii) Doubtful Conclusions: This method may lead to doubtful conclusions due to carelessness of the witnesses.

Difference between Direct Personal Investigation and Indirect Oral Investigation

The difference between direct personal investigation and indirect oral investigation is as under:

(i) In the case of direct personal investigation, the investigator establishes direct contact with the informants. On the other hand, in the case of indirect oral investigation, information is obtained by contacting other than those about whom information is sought.

(ii) Direct Personal Investigation is generally possible when the field of investigation is small. On the other hand, indirect oral investigation is generally preferred when the field of investigation is relatively large.

(iii) In the Direct Personal Investigation, the investigator must be well versed in the language and cultural habits of the informants. There is no such requirement in the case of Indirect Oral Investigation.

(iv) Direct investigation is relatively costlier than the indirect investigation.

(3) Information from Local Sources or Correspondents

Under this method, the investigator appoints local persons or correspondents at different places. They collect information in their own way and furnish the same to the investigator.

Suitability

This method is suitable particularly when:

(i) regular and continuous information is needed.

(ii) the area of investigation is large.

(iii) the information is to be used by journals, magazines, radio, TV, etc. and

(iv) a very high degree of accuracy of information is not required.

Merits

Principal merits of this method are as under:

(i) Economical: This method is quite economical in terms of time, money or efforts involved.

(ii) Wide Coverage: This method allows a fairly wide coverage of investigation.

(iii) Continuity: The correspondents keep on supplying almost regular information.

(iv) Suitable for Special Purpose: This method is particularly suitable for some special-purpose investigations, e.g., price quotations from the different grain markets for the construction of Index Number of agricultural prices.

Demerits

Following are some notable demerits of this method:

(i) Loss of Originality: Originality of data is sacrificed owing to the lack of personal contact with the respondents.

(ii) Lack of Uniformity: There is lack of uniformity of data. This is because data is collected by a number of correspondents.

(iii) Personal Bias: This method suffers from the personal bias of the correspondents.

(iv) Less Accurate: The data collected by this method are not very accurate.

(v) Delay in Collection: Generally, there is a delay in the collection of information through this method.

(4) Information through Questionnaires and Schedules

Under this method, the investigator prepares a questionnaire keeping in view the objective of the enquiry. There are two ways of collecting information on the basis of questionnaire:

(i) Mailing Method, and

(ii) Enumerator’s Method.

(i) Mailing Method

Under this method, questionnaires are mailed to the informants. A letter is attached with the questionnaire giving the purpose of enquiry. It is also assured that the information would be kept secret. The informant notes the answers against the questions and returns the completed questionnaire to the investigator.

Suitability

This method is most suited when:

(a) the area of the study is very wide, and

(b) the informants are educated.

Merits

The following are the main merits of this method:

(a) Economical: This method is economical in terms of time, money and efforts involved.

(b) Original: This method is original and therefore, fairly reliable. This is because the information is duly supplied by the concerned persons themselves.

(c) Wide Coverage: This method allows wide coverage of the area of study.

Demerits

However, there are certain demerits of this method as under:

(a) Lack of Interest: Generally, the informants do not take interest in questionnaires and fail to return the questionnaires. Those who return, often send incomplete answers.

(b) Lack of Flexibility: This method lacks flexibility. When questions are not properly replied, these cannot be changed to obtain the required information.

(c) Limited Use: This method has limited use owing to the fact that the questionnaires can be answered only by the educated informants. Thus, this method cannot be used when the informants are uneducated.

(d) Biased: If the informants are biased, then the information will also be biased.

(e) Less Accuracy: The conclusions based on such investigation have only limited accuracy. This is because some questions may be difficult, and consequently accurate answers may not be offered.

(ii) Enumerator's Method

Under this method, a questionnaire is prepared according to the purpose of enquiry. The enumerator himself approaches the informant with the questionnaire. The questionnaires which are filled by the enumerators themselves by putting questions are called schedules. Thus, under this method, the enumerator himself fills the schedules after seeking information from the informants. Enumerators are those persons who help the investigators in collecting the data. The enumerators are given training to fill the schedules and put the questions intelligently to obtain accurate information.

Suitability

This method is mostly used when:

(a) field of investigation is large.

(b) the investigation needs specialised and skilled investigators, and

(c) the investigators are well versed in the local language and cultural norms of the informants.

Investigator, Enumerator and Respondent

- Investigator is a person who plans and conducts an empirical investigation independently or with the help of others.
- Enumerator is a person who actually collects the desired statistical information or statistical data. Often the enumerators are trained personnel hired by the investigator for field work.
- Respondent is a person who answers/responds to the set of questions included in the questionnaire.

Merits

This method has the following merits:

(a) Wide Coverage: This method is capable of a wide coverage in terms of the area involved. Even illiterates can furnish the required information.

(b) Accuracy: There is a fair degree of accuracy in the results. This is because investigations are done by specialized enumerators.

(c) **Personal Contact:** Unlike in the case of mailing questionnaires, there is personal contact with the informants in this method. Accordingly, accurate and right answers are obtained.

(d) Impartiality: This method is impartial. This is because the enumerators themselves do not need the required information, so they are impartial to the nature of information that they obtain.

(e) Completeness: Schedules have the merit of completeness because these are tilled in by the enumerators themselves.

Demerits

The following are the main demerits of this method:

(a) Expensive: This is a very expensive method of investigation because of the involvement of trained investigators.

(b) Availability of Enumerators: Competent enumerators may not be available. Accuracy of the information accordingly suffers.

(c) Time Consuming: Enumerators may need specialised training for particular investigation. The process of investigation thus becomes time consuming.

(d) Not Suitable for Private Investigation: Since this method is very expensive, it is generally not suitable for private investigation. This method is generally used by the Government institutions.

(e) **Partial:** If the enumerators are biased, then the data will not be accurate.

Construction of Questionnaires and Schedules and their Qualities

In the context of collection of Primary data, construction of questionnaires and

schedules has the special significance.

The Basic Difference between a Questionnaire and a Schedule

Both show a set of questions. But in case of a questionnaire, the answers are to be recorded/written by the respondents themselves, while in the case of a schedule, answers are to be written/recorded by the enumerators specifically hired for the purpose.

The set of questions in questionnaires and schedules are similar. The only difference that lies between the two is that, in questionnaires, the entire information is recorded by the informants themselves. In the schedules, on the other hand, the information as supplied by the informants is recorded by the enumerators.

Qualities of a Good Questionnaire

Following are some of the desired qualities of a good questionnaire:

(1) Limited Number of Questions: The number of questions in a questionnaire should be as limited as possible. Questions should be only relating to the purpose of enquiry.

(2) Simplicity: Language of the questions should be simple, lucid and clear. Questions should be short, not long or complex. Mathematical questions must be avoided.

(3) Proper Order of the Questions: Questions must be placed in a proper order.

(4) No Undesirable Questions: Undesirable questions or personal questions must be avoided. The questions should not offend the informants.

(5) **Non-Controversial:** Questions should be such as can be answered impartially. No controversial questions should be asked.

(6) Calculations: Questions involving calculations by the respondents must be avoided. Investigator himself should do the calculation job.

(7) **Pre-Testing Pilot Survey:** Some questions be asked from the informants on trial basis. If their answers involve some difficulty these can be reframed accordingly. Such testing is technically called pilot survey.

(8) **Instructions:** A questionnaire must show clear instructions for filling in the form.

(9) **Cross Verification:** Such questions may be asked which help cross verifications.

(10) Request for Return: Request should be made to the respondents to return the questionnaire completed in all respects. The informant must be assured that the information conveyed by him will be treated as confidential.

Types of Questions: Some Examples

There are four possible types of questions, as under:

(1) Simple Alternative Questions: These questions are answered in ‘Yes’ or ‘No’, ‘Right or Wrong’ and ‘Good or Bad.’

Example

Do you have a car?

Or

Yes/No

Government decides to introduce 10+2 system in the college.

Do you agree?

Yes/No

(2) Multiple Choice Questions: When there are various possibilities of a particular event, multiple choice questions are generally asked. A number of possible answers are given for such questions. The informant is to tick-mark the one that he feels fit.

Example

What is your mode of conveyance from home to college?

1. On foot
2. Cycle
3. Bus
4. Scooter

Correct answer be tick-marked (/) in the box.

(3) **Specific Information Questions:** Only specific information is obtained through such questions.

Example

In which class do you read?

Or

How much pocket allowance do you get?

(4) Open Questions: In such questions, the informant is requested to give his views on specific issues.

Example

How can prices in India be controlled?

Or

How can power shortage in the country be overcome?

Example of an Ideal Questionnaire

Objective of this questionnaire is to know about the monthly income and expenditure of the 10+2 students living in the hostels. You are requested to fill in this questionnaire, and return at your earliest convenience. Information furnished in this questionnaire will be kept strictly confidential. The information will be used only for the present investigation.

1. Student’s Name
2. Age
3. Faculty.................................. Art/Commerce/Science
4. Name of the School I College
5. Father’s name and address

6. Father's Occupation.................................Income

7. Income (if any) of other members in the family

8. Monthly income received by the student

(i) From the family ....

(ii) Personal earning..

(iii) Scholarship ....

(iv) Others ....

9. Monthly Expenditure of the Student

Items of Expenditure	Amount of Expenditure
(i) School/College fee	....
(ii) Stationery	....
(iii) Books	....
(iv) Conveyance	....
(v) Hostel expenses	....
(Vi) Entertainment	....
(vii) Other items (specify)	....

10. Is your monthly income enough for you? Yes ....No

11. If your monthly income is not enough, how do you propose to increase it?

12. Can you save anything from your monthly income? If yes, under which of the above-noted heads of expenditures can you save and how much?

Principal Steps in the Planning of a Field Survey (or Field Investigation)

While planning a field survey, you are required to focus on the following steps:

(i) Design questionnaire with utmost care and be sure that:

(a) the questionnaire has a reasonable length.

(b) the questionnaire includes only precise and short questions.

(c) the set of questions in the questionnaire can be cross-checked with each other.

(d) the questions do not involve seriousl difficult calculations for the respondent.

(ii) Decide the mode of enquiry, viz. direct personal oral investigation (also called interview method) or mailing the questionnaire.

(iii) Arrange a proper training programme for the enumerators, explaining to them the purpose and mode of enquiry and also the nature of various questions in the questionnaire.

(iv) Conduct a pilot survey (a small preliminary investigation) when the field of investigation is very large.

Pilot Survey

What it is? It refers to a try-out survey covering a very small sample of the universe of your study. Why do it? This is a sort of pre-testing of your questionnaire. This helps you to assess quality of your questionnaire and the way respondents respond to the set of questions. Accordingly, you are able to know in advance the shortcomings/drawbacks of your questionnaire. Required changes in the questionnaire may be introduced before you are set out for a final survey.

Pilot Survey helps:

(i) in assessing the quality and suitability of questions

(ii) in assessing performance of the enumerators

(iii) in designing a set of instructions for the enumerators

(iv) in assessing the cost and time involved in the final survey.

Main Sources of Errors in Collection of Data

These are the following:

(i) Errors related to the measurement of objects which may occur when: (a) the scale of measurement happens to be different for different enumerators, and (b) different enumerators may be allowing different degree of approximation in their measurement, even while using identical scales.

(ii) Errors occurring due to wrong responses simply because the respondents are not able to handle/understand the questions precisely.

(Hi) Errors occurring due to the lack of response. Some respondents may not respond to the questionnaire. Lack of information thus occurring infuses an element of error in the collection of data.

(iv) Errors occurring due to miscalculations, also called arithmetical errors.

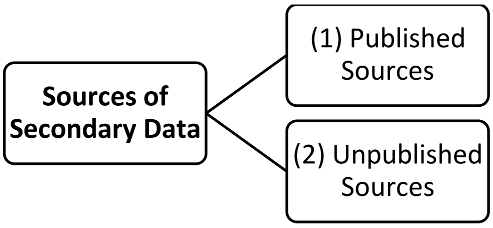
(v) Errors occurring due to 'communication gap 'or due to lack of recording of the information.

3. COLLECTION OF SECONDARY DATA

There are two main sources of secondary data:

(1) Published Sources, and

(2) Unpublished Sources.



(1) Published Sources

Some of the published sources of secondary data are:

(i) Government Publications: Ministries of the Central and State Governments in India publish a variety of Statistics as their routine activity. As these are published by the Government, data are fairly reliable. Some of the notable Government publications on Statistics are: Statistical Abstract of India, Annual Survey of Industries, Agricultural

Statistics of India, Report on Currency and Banking, Labour Gazette, Reserve Bank of India Bulletin, etc.

(ii) Semi-Government Publications: Semi-Government bodies (such as Municipalities and Metropolitan Councils) publish data relating to education, health, births and deaths. These data are also fairly reliable and useful.

(iii) Reports of Committees and Commissions: Committees and Commissions appointed by the Government also furnish a lot of statistical information in their reports. Finance Commission, Monopolies Commission, Planning Commission are some of the notable commissions in India which supply detailed statistical information in their reports.

(iv) Publications of Trade Associations: Some of the big trade associations, through their statistical and research divisions, collect and publish data on various aspects of trading activity. For example, Sugar Mills Association publishes information regarding sugar mills in India.

(v) Publications of Research Institutions: Various universities and research institutions publish information as findings of their research activities. In India, for example, Indian Statistical Institute, National Council of Applied Economic Research publish a variety of statistical data as a regular feature.

(vi) Journals and Papers: Many newspapers such as **‘The Economic Times’** as well as magazines such as **Commerce, Facts for You** also supply a large variety of statistical information.

(vii) Publications of Research Scholars: Individual research scholars also sometimes publish their research work containing some useful statistical information.

(viii) International Publications: International organisations such as UNO, IMF, World Bank, ILO, and foreign governments etc., also publish a lot of statistical information. These are used as secondary data.

(2) Unpublished Sources

There are some unpublished secondary data as well. These data are collected by the government organisations and others, generally for their self use or office record. These data are not published. This unpublished numerical information may, however, be used as secondary data.

A Note of Caution for the Users of Secondary Data

Users of secondary data must check:

(i) reliability of data,

(ii) suitability of data, and

(iii) adequacy of data.

Precautions in the Use of Secondary Data

We know that secondary data are collected by others to suit, their specific requirements. Therefore, one needs to be very careful while using these data. **Connor** has rightly stated, “Statistics especially other people’s Statistics are full of pitfalls for the users.” Some of the notable questions to be borne in mind while dealing with the secondary data are:

(i) Whether the data are reliable?

(ii) Whether the data are suitable for the purpose of enquiry?

(iii) Whether the data are adequate?

In order to assess the reliability, suitability and adequacy of the data, the following points must be kept in mind:

(1) Ability of the Collecting Organisation: One should check the ability of the organisation which initially collected the data. The data should be used only if it is collected by able, experienced and impartial investigators.

(2) Objective and Scope: One should note the objective of collecting data as well as the scope of investigation. Data should be used only if the objective and scope of the study as undertaken earlier match with the objective and scope of the present study.

(3) **Method of Collection:** The method of collection of data by the original investigator should also be noted. The method adopted must match the nature of investigation.

(4) **Time and Conditions of Collection:** One should also make sure of the period of investigation as well as the conditions of investigations. For example, data collected during war times may not be suitable to generalise certain facts during peace times.

(5) **Definition of the Unit:** One should also make sure that the units of measurement used in the initial collection of data are the same as adopted in the present study. If the unit of measurement differs, data must be modified before use.

(6) **Accuracy:** Accuracy of the data should also be checked. If the available data do not conform to the required degree of accuracy, these should be discarded.

In short, as stated by **Bowley,** “It is never safe to take published Statistics at their face value without knowing their meaning and limitations”

Two important Sources of Secondary Data: 'Census of India' and Reports and Publications of National Sample Survey Office'

(1) Census of India: Census of India is a decennial publication of the Government of India. It is published by Registrar General & Census Commissioner, India. It is a very comprehensive source of secondary data. It relates to population size and the various aspects of demographic changes in India. Broadly, it includes statistical information on the following parameters:

(i) Size, growth rate and distribution of population in India.

(ii) Population projections.

(iii) Density of population.

(iv) Sex composition of population.

(v) State of literacy.

Information on these parameters relates to country as a whole as well as different states and union territories of the country. As the name suggests, Census of India is a comprehensive enquiry on population size and the related parameters of change covering each and every household of the country.

(2) Reports and Publications of National Sample Survey Office (NSSO): Reports and publications of NSSO is another important source of secondary data in India. NSSO is a government organisation under the Ministry of Statistics and Programme Implementation. This organisation conducts regular sample surveys to collect basic statistical information relating to a variety of economic activity in rural as well as urban parts of the country. For example, the 76th round of NSSO (July 2018-December 2018) was on “Persons with Disabilities, and Drinking Water, Sanitation, Hygiene and Housing Conditions”. Broadly, reports and publications of NSSO offers statistical information of the following parameters of economic change:

important

The statistical data collected by NSSO are released through its quarterly journal, called SARVEKSHANA and its reports, popularly known as NSSO Reports.

(i) Land and Livestock Holdings.

(ii) Housing Conditions and Migration with special emphasis on slum dwellers.

(iii) Employment and Unemployment status in India.

(iv) Consumer Expenditure in India, including level and pattern of consumer expenditure of diverse categories of the people.

(v) Sources of Household income in India.

Unlike Census of India, Reports and Publications of National Sample Survey Office are based on ‘sample’ study of the population/universe.

Important Agencies at the national level which collect, process and tabulate the statistical data: NSSO (National Sample Survey Office), RGI (Registrar General of India), DGCIS (Directorate General of Commercial Intelligence and Statistics) and Labour Bureau.

Multiple Choice Questions

Select the correct alternative:

1. Data collected for the first time from the source of origin is called:

(a) primary data

(b) secondary data

(c) internal data

(d) none of these

2. What kind of data are contained in the census of population and national income estimates, for the government?

(a) Primary data

(b) Secondary data

(c) Internal data

(d) None of these

3. Which of the following is a method of secondary data collection?

(a) Direct personal investigation

(b) Direct oral investigation

(c) Collection of information through questionnaire

(d) None of these

4. Which of the following is a merit of a good questionnaire?

(a) Difficulty

(b) Less number of questions

(c) Not in proper order

(d) Invalid questions

5. Which of the following methods is used when an investigator collects the required information with the informant?

(a) Direct Personal Investigation

(b) Indirect Oral Investigation

(c) Mailing Method

(d) Enumerator's Method

6. In order to know the likings and dislikings of the listeners of the programmes broadcast by the Himachal Akashvani, the latter is keen to collect data. Which method of collecting data will be suitable for it?

(a) Direct Personal Investigation

(b) Indirect Oral Investigation

(c) Mailing Method

(d) Enumerator's Method

7. Schedules are filled by the:

(a) investigator

(b) enumerator

(c) informant

(d) none of these

8. Which of the following is a source of secondary data?

(a) Government publication

(b) Private publication

(c) Report published by the State Bank of India

(d) All of these

9. Which data is collected by the investigator himself`

(a) Primary

(b) Secondary

(c) Both (a) and (b) above

(d) Neither of the above

10. Data collected from The Times of India’ is an example of:

(a) Primary Data

(b) Secondary Data

(C) None of these

(d) Census

11. Data collected on religion from the census reports are:

(a) Sample data

(b) Secondary data

(C) Primary data

(d) Either (a) or (b)

12. When population under investigation is infinite, we should use:

(a) Sample method

(b) Census method

(C) Either census or sample method

(d) Neither census nor sample method

13. Data from secondary source are:

(a) Collected for other purposes than the current study

(b) Obtained from the newspaper

(c) More reliable than data from a primary source

(d) Both (a) and (b)

14. Primary data is preferred over secondary data where:

(a) Time available is short

(b) Accuracy is important

(c) Sufficient finance is not available

(d) Much accuracy is not required

15. The data collected on the height of a group of students after recording their heights with a measuring tape are:

(a) Primary data

(b) Continuous data

(c) Discrete data

(d) Secondary data

16. After every ten years, information regarding population of India is collected through:

(a) Census

(b) Sample

(c) Both (a) and (b)

(d) Neither of the above

17. Sample study is useful:

(a) When population is not completely known

(b) As it is easy to handle samples

(C) As results are more reliable

(d) It is cheap

18. Which method of collection of data covers the widest area`

(a) Direct Personal Investigation

(b)Mailed Questionnaire Method,

(c) Telephone interview method

(d) All of these

19. It is best to use a census while conducting a survey if:

(a) The population is large

(b) The population is small

(c) Time is limited to conduct the survey

(d) Cheaper method is needed

20. A good questionnaire should have:

(a) Minimum questions

(b) Concise

(c) Clear

(d) All the above

21. In random sampling:

(a) Each element has equal chance of being selected

(b) Sample is always full of bias

(c) Cost involved is very less

(d) Cost involved is high

22. Direct personal investigation method suffers from:

(a) Personal Bias

(b) Excessive expenses

(c) Time Consuming

(d) All the above

23. The quickest method to collect primary data is:

(a) Direct Personal Investigation

(b) Indirect Oral Investigation

(c) Telephone Interview

(d) Mailed Questionnaire Method

24. Stratified sample is preferred where:

(a) Population is perfectly homogeneous

(b) Population is non-homogeneous

(c) Random sampling is not possible

(d) Small samples are required

25. Which of the following methods is used for the estimation of population in country`

(a) Sampling Method

(b) Census Method

(c) Both (a) and (b)

(d) Neither (a) nor (b)

26. Data collected on religion from the census reports are:

(a) Secondary Data

(b) Primary Data

(c) Sample Data

(d) Either (a) or (b)

27. Which method of collection of data covers the widest area`

(a) Telephone interview method

(b) Mailed questionnaire method

(c) Direct interview method

(d) All of these

Answers

Multiple Choice Questions

1. (a)	**2. (b)**	**3. (d)**	**4. (b)**	**5. (a)**
6. (d)	**7. (b)**	**8. (a)**	**9. (a)**	**10. (b)**
11. (b)	**12. (a)**	**13. (d)**	**14. (b)**	**15. (a)**
16. (a)	**17. (a)**	**18. (b)**	**19. (b)**	**20. (d)**
21. (a)	**22. (d)**	**23. (c)**	**24. (b)**	**25. (b)**
26. (a)	**27. (b)**			

CHAPTER 3

CENSUS AND SAMPLE METHODS OF COLLECTION OF DATA

There are 2,000 students in a college. An investigator wants to collect data regarding their family background. He has two possible choices. First, he collects information relating to all the

2.000 students. Second, he collects information relating to some of the students (sample of students) who would represent all the

2.000 students. In Statistics, the first approach for collecting data is called **Census Method** and the second approach is called **Sample Method.** The present chapter focuses on a comprehensive study of the census and sample methods of data collection. Beginning with the concepts of Census and Sample techniques the chapter offers a detailed discussion of the merits and demerits of the two techniques. Also, various techniques of 'sampling' are discussed with a comparative look at their merits and demerits.

1. CONCEPTS OF 'CENSUS' AND 'SAMPLE'

Chapter 2 of the book introduces the concept of universe or population. It would be useful to recapitulate this concept for a comprehensive understanding of the concepts of 'Census' and 'Sample'.

In Statistics, universe or population simply refers to an aggregate of items to be studied for an investigation. Ordinarily, the term population is used to mean total number of people living in a country. Population of India was approximately 121.02 crore in 2010-11. But in Statistics, the term population is used differently. In Statistics, the term population means the aggregate of all items about which we want to obtain information. To illustrate, there are 2,000 students in a particular college. If an investigation relates to all the 2,000 students, then 2,000 would be taken as universe or population. Each unit of these 2,000 is called Item. To further illustrate, I sugar mill out of the 10 sugar mills we are studying, would be called an item. All the 10 sugar mills would constitute population or the universe.

If a statistical inquiry is based on all items of the universe, it is called a census inquiry. For example, if you want to know quality of life of the 25,000 households in your town and you decide to collect the relevant statistical data of all 25,000 households (that is, your statistical inquiry is covering all the items of the universe or is covering the entire universe) you are relying on census method of your statistical inquiry. Alternative is that you collect statistical data for every 5th or 10th household of your town, which you think should represent all the 25,000 households of the town. Now you are not covering each and every item of the universe; instead you are covering only a 'sample' of the universe. Characteristics of the 'sample' are supposed to represent characteristics of the entire universe. Or, quality of life of a sample of (say) 2,500 families is supposed to represent quality of life of all the 25,000 families of your town. This is called a sample method of a statistical inquiry.

The concepts of 'Sample' and 'Population' are projected in Fig. 1. Bigger part of this picture is comprising the entire 27 items showing population or universe and the corresponding part comprising of 3 items shows sample.

Now you must be sure that Sample is only a part of the population or the universe. But it must be that part which, in terms of its characteristics, represents the entire population.

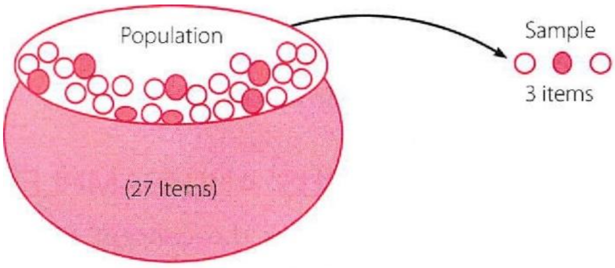
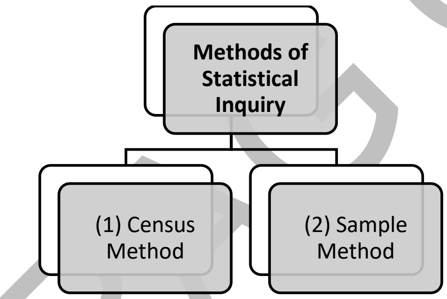


Fig. 1



2. CENSUS METHOD

Census method is that method in which data are collected covering every item of the universe or population relating to the problem under investigation.

To illustrate, you may be interested in the investigation of colour composition of the Maruti cars in India. According to the Census Method, you are required to collect data on the colour of each and every Maruti car sold in India.

Census method implies complete enumeration of the universe/ population. Census of population is the most suitable example of tbe census method of a statistical enquiry. For the estimation of the country's population, house to house enquiry is conducted and even people living at the roadside are contacted in India, census of population is conducted every ten years, and the last census was conducted in February 2011.

Census of India 20V reveals that in terms of the size of population. India is the second largest country in the world next only to China.

Suitability

Census method is suitable particularly for such statistical investigations which have (i) small size of population, (ii) widely diverse items in the population, (iii) requirement of intensive examination of different items, and (iv) high degree of accuracy and reliability.

Merits

Principal merits of census method are as under:

(1) Reliable and Accurate: Results based on census method are accurate and highly reliable. This is because each and every item of the population is studied.

(2) Less Biased: Results based on census method are less biased. It is because of the absence of investigator's discretion regarding the selection of sample items.

(3) **Extensive Information:** Information collected through the census method is quite exhaustive and therefore, more meaningful because all the items of a universe are examined. For example, population census in India gives exhaustive information relating to the number of people in different parts of the country, their age and sex composition, education, status, occupation, and the like.

(4) **Study of Diverse Characteristics:** By using census method, one can study diverse characteristics of the universe.

(5) Study of Complex Investigation: When items in a universe are of complex nature and it is necessary to study each item, only census method can produce the desired results. Data on country's population are collected by this method.

(6) Indirect Investigation: Census method can be successfully used in indirect investigations relating to unemployment, poverty, corruption, etc.

Demerits

However, there are certain demerits of census method as under:

(1) Costly: Census method is very costly and is, therefore, generally not used for ordinary investigations. Only the Government or some big institutions can afford to use this method and that too for specific purposes only.

(2) Large Manpower: Census method requires large manpower (enumerators). Training of a large number of enumerators becomes essential, which is a very difficult process.

(3) **Not Suitable for Large Investigations:** If the universe comprises a large number of items, then it may not be possible to cover each and every item. Census method becomes practically inoperative in such situations.

3. SAMPLE METHOD

Sample method is that method in which data is collected about the sample on a group of items taken from the population for examination and conclusions are drawn on their basis.

Sample method is widely used in our day-to-day life. A lady in the kitchen, for example, tests only a grain or two of the rice to know whether the rice is boiled or not. By examining only a few drops of blood, a doctor determines the blood group of a person.

Suitability

Sample method is particularly suitable when: (i) the size of population is very large, (ii) very high degree of accuracy is not needed, (iii) intensive examination of diverse items is not required, and (iv) when different units of the universe are broadly similar to each other.

Merits

Some of the principal merits of the sample method are as under:

(1) Economical: Sample method of investigation is economical because only some units of the population are studied.

(2) Time Saving: In this method, only limited number of the items are investigated. As such the process of investigation is time-saving, not time-consuming.

(3) **Identification of Error:** Because only a limited number of items are covered, errors can be easily identified. To that extent sampling method shows better accuracy.

(4) Large Investigations: Sample method is more feasible in situations of large investigations than the census method which generally involves unaffordable cost.

(5) **Administrative Convenience:** There is an administrative convenience in handling a limited number of items. More capable and efficient investigators can be appointed.

(6) More Scientific: According to **R. Fisher,** Sample Method is more scientific because the sample data can be conveniently- investigated from various angles.

Demerits

Yet there are some demerits of the sample method as under:

(1) Partial: It is only a partial investigation of the universe. The investigator’s bias in the selection of the sample is not ruled out. Accordingly, the results may be biased as well.

(2) **Wrong Conclusions:** If the selected sample does not represent the characteristics of the universe, the study may end up with wrong conclusions.

(3) Difficulty in Selecting Representative Sample: It is not very easy to select a sample which would represent the characteristics of the entire population.

(4) **Difficulty in Framing a Sample:** Sometimes the universe may be so diverse that it becomes difficult to frame a sample.

(5) **Specialised Knowledge:** Sampling involves a set of technical procedures. One must have the technical knowledge of choosing a representative sample from the universe. Persons who are well-versed with all the techniques of sampling are not easily available.

Two Basic Essentials of a Good Sample

These are:

(i) that the sample must represent characteristics of the entire universe/ population.

(ii) that the sample must be adequate enough to offer reliable conclusions.

Essentials of a Sample

In order to arrive at an unbiased and right conclusion, a sample must have the following qualities or essentials:

(1) Representative: A sample must represent all the characteristics of the universe. It is possible only when each unit of the universe stands equal chances of being selected in the sample.

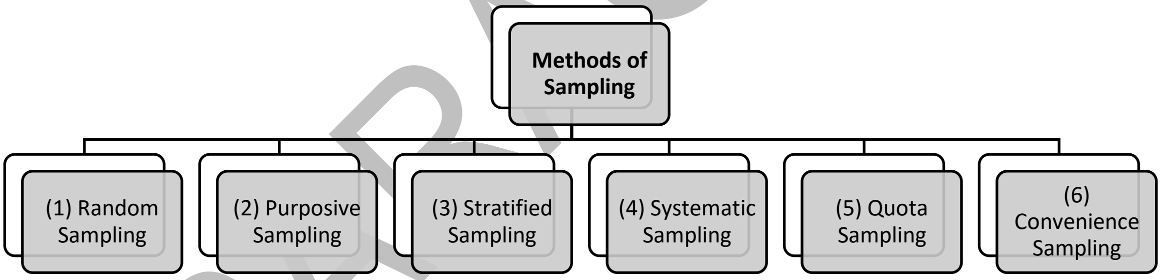
(2) **Independent:** All units of a sample must be independent of each other. In other words, inclusion of one item in tire sample should not be dependent upon the inclusion of some other items of the universe.

(3) **Homogeneity:** If more than one sample are selected from a universe, these samples should be homogeneous (and not contradictory) to each other.

(4) Adequacy: The number of items in the sample should be fairly adequate so that some reliable conclusions are drawn covering the characteristics of the universe as a whole.

METHODS OF SAMPLING

Following are the principal methods or techniques of sampling:



(1) Random Sampling

Random sampling is that method of sampling in which each and every item of the universe has equal chance of being selected in the sample. In other words, there is an equal probability for every item of the universe being selected in the sample.

Which of the various items of the universe would get selected in the sample is beyond the control of the investigator. The selection is left entirely to the chance factors. This method is used particularly when various items of the universe are homogeneous or identical to each other. This method is impartial and economical. Random Sampling may be done in any of the following ways:

(i) Lottery Method: In this method, paper-slips are made for each item of the universe. These slips are shuffled in a box. Then, impartially, some of the slips are drawn to form a sample of the universe.

(ii) Tables of Random Numbers: Some statisticians have prepared a set of tables called Tables of Random Numbers. A sample is framed with reference to these tables. Of all these tables Tippet's Table is most widely used. Using 41,600 figures, Tippet has involved 10,400 numbers comprising of four units each. For the use of this method, all items of the universe are first arranged in an order. Then using Tippet's Table the required number of items are selected as are needed for a sample.

Merits

Following points may be noted on the merits of random sampling:

(i) This method is free from personal bias of the investigator.

(ii) Each and every item of the universe stands equal chances of being selected.

(iii) The universe gets fairly represented by the sample.

(iv) This is a very simple and straightforward method.

The Principal Merit of Random Sampling

It is that each and every item of the universe has equal chance (or equal probability) of being selected.

Demerits

However, there are two notable demerits of this method. These are as under:

(i) This method does not guarantee proportionate representation of different items in the universe.

(ii) Random sampling does not give weightage to certain important items in the universe.

Difference Between Random sampling and Haphazard sampling

Random Sampling	Haphazard Sampling
(i) Random sampling is in accordance with the rules of sampling.	(i) Haphazard sampling is not in accordance with the rules of sampling.
(ii) Random sampling allows every item an equal chance (or equal probability) of being selected in the sample.	(ii) Haphazard sampling does not allow every item an equal chance (or equal probability) of being selected in the sample.

Random Sampling and Exit Polls

Exit polls is an interesting example of random sampling. What does it mean? It means a small percentage of the people exiting the poll booths are contacted and asked whom they voted for. From the sample of information thus collected, a prediction is made about the victory chances of different candidates contesting election.

(2) Purposive or Deliberate Sampling

Purposive sampling is that method in which the investigator himself makes the choice of the sample items which in his opinion are the best representative of the universe. Thus, in this method of sampling, selection of the sample items is not left to the chance factors; it is simply made by choice.

This method of sampling is specifically suitable when some of the items in the universe are of special significance and ought to be included in the sample. For example, if an investigation is to be made relating to the iron and steel industry in India, then the inclusion of such industries as the Tata Iron and Steel Company is obviously a purposive selection.

However, there is a considerable possibility of personal bias in purposive sampling. As a result, it loses its credibility.

Merits

(i) This method is flexible to allow the inclusion of those items in the sample which are of special significance.

(ii) Selection of items can be deliberately tuned to the purpose of study.

(iii) It is a very simple technique of selection of the sample items.

Demerits

(i) There is a possibility of personal bias in the selection of items.

(ii) Because of the possibilities of personal bias, reliability of the results becomes doubtful.

(3) Stratified or Mixed Sampling

This method of sampling is generally adopted when population consists of different groups with different characteristics. According to this method of sampling, population is divided into different strata having different characteristics and some of the items are selected from each strata, so that the entire population gets represented. Each stratum should be represented in correct proportion in the sample. To illustrate, suppose there are 50 students in Class XL Out of them, 30 have studied Mathematics and 20 have studied Home Science in their Class X. Thus, the population of the 50 students gets divided into two strata consisting of 30 and 20 students respectively. From each of these strata, items would be selected proportionately such that the sample represents the characteristics of the entire population. If, of the total 50 students, only a sample of five is to be selected, then we shall randomly select three out of the first stratum (comprising 30 students) and two out of the second stratum (comprising 20 students). However, items may also be selected unproportionately from different strata.

An investigator may prefer to select four students from amongst those who studied Mathematics and only one from amongst those who studied Home Science. The choice will be governed by the nature of the enquiry and wisdom of the investigator.

Stratified Sampling is also called Mixed technique of sampling because this method involves the mixture of both purposive sampling and random sampling. The division of population into different strata is purposely done while selection of the items is done at random.

Merits

(i) This method covers diverse characteristics of the population.

(ii) On the basis of diverse characteristics of the population, a comparative analysis of the data becomes possible.

(iii) This method of sampling offers reliable as well as meaningful results.

Demerits

(i) This method is suitable only when there is a complete knowledge about the diverse characteristics of the population. Therefore, this has a limited scope.

(ii) There is a possibility of bias at the time of classification of the population into different strata.

(iii) When the size of population is already small, it may be difficult to further divide it into smaller parts/strata.

(4) Systematic Sampling

According to this method, units of the population are numerically, geographically and alphabetically arranged. Every nth item of the numbered items is selected as a sample item. To illustrate, if 10 out of 100 students are to be selected for a sample, then 100 students would be numbered and systematically arranged. One item of the first 10 would be selected at random. Subsequently, every 10th item from the selected number will be selected to frame a sample. If the first selected number is 5th item, then the subsequent numbers would be 15th, 25th, 35th, 45th, 55th, 65th, 75th, 85th and 95th. This method of sampling is, in fact, a short-cut method of Random Sampling.

Merits

(i) This is a very simple method. Sample is easily determined.

(ii) There is hardly any possibility of personal bias in this method.

The Principal Merit of Stratified Sampling

It allows selection of such items which represent diverse characteristics of the population.

Principal Demerit of Systematic Sampling

It does not allow every item of the universe equal chance (or equal probability) of being selected in the sample.

Demerits

(i) Every item in the population does not get equal chance of being selected because only the first item is selected on the basis of random sampling.

(ii) If all the items in the population are homogeneous, this method of sampling serves no specific purpose.

(5) Quota Sampling

In this method, the population is divided into different groups or classes according to different characteristics of the population. Some percentage of different groups in total population is fixed. Further, some quota of the items to be selected as sample-items is fixed for each group. The investigator selects the fixed number of items from each group to frame a sample.

This method of sampling is not very expensive. But there is a high possibility of personal bias at the time of selection of the items by the investigator. Accordingly, the reliability of results becomes questionable.

(6) Convenience Sampling

In this method, sampling is done by the. investigator in such a manner that suits his convenience. To illustrate, an investigator may select a sample of teachers merely by referring to the college prospectus. This method is the simplest and least expensive, but unscientific and unreliable. It depends too much on the whims of enumerators.

Reliability of Sampling Data

The reliability of the sampling data means that the characteristics of the universe are fully represented by the sample. It depends mainly on the following factors:

(1) Size of the Sample: Reliability of sampling depends on the size of the sample. If its size is very small, it will fail to represent the population. Accordingly, the conclusions would lack reliability.

(2) Method of Sampling: If the method of sampling is not simple and exhaustive, it will not adequately represent the population. Consequently, the results will not be dependable.

(3) Bias of Correspondents and Enumerators: Personal bias of the correspondents and enumerators should be as less as possible. Otherwise, reliability of the sampling data is bound to suffer.

(4) Training of Enumerators: Reliability of sample also depends upon the training of the investigators. If they are not trained to make them expert in their field of investigation, the sample will lack reliability.

Census and Sampling Methods: A Comparative Look

Some of the principal differences between census and sampling methods are as under:

(1) Coverage: In the census technique, an investigator collects information relating to all the items in the population. In sampling method, on the other hand, only some of the items which represent the population are covered for an investigation.

(2) Suitability: Census method is suitable when the area of investigation is relatively small. On the other hand, when the area of investigation is large, it is the sampling method which is generally used.

(3) **Accuracy:** There is generally a greater degree of accuracy in the results based on the census method than the sampling method. This is because in the census method each and every item of the population is studied. As against it, there is less accuracy and reliability in the sampling method because it studies only a few items of population. However, errors can be easily detected and removed in the sampling method because of the small number of items. To that extent sampling method shows greater accuracy than the census method.

(4) Cost: Sampling method is certainly much less expensive than the census method. Smaller the sample size out of the given population, lesser the cost of investigation.

(5) Time: Sample method is less time consuming than the census method.

(6) Nature of Items: Census method is particularly suitable when the items in the population have diverse characteristics. On the other hand, sample method is suitable when items in the population are homogeneous.

(7) Verification: Verification of the statistical information obtained through census method is generally not possible. It would involve huge expenses and the repetition of the whole process. Sample information, on the other hand, can be easily verified. In case of doubt, enumeration can be done again and facts verified accordingly.

In short, sample method of statistical investigation is generally preferred to the census method because the former is less expensive in terms of the time, money and efforts involved. However, for the successful application of the sample method, it is very essential that the sample items represent the characteristics of population as a whole.

Statistical Errors: Sampling and Non-Sampling Errors

Statistical errors are broadly classified as (i) sampling errors, and (ii) non-sampling errors. Following are the details:

(i) Sampling Errors: These are related to the size or nature of the sample selected for the study. Due to a very small size of the sample selected for study or due to non-representative nature of the sample, the estimated value may differ from the actual value of a parameter. The error thus emerging, is called sampling error. For example, if the estimated value of a parameter is found to be 10 while the actual/true value is 20 then, the sampling error = estimated value of the parameter - true value of the parameter = 10-20 = -10.

(ii) Non-sampling Errors: These are errors related to the collection of data. These are of the following types:

Error of Measurement: Error of measurement may occur due to.- (a) difference in the scale of measurement, and (b) difference in the rounding off procedure adopted by different investigators.

Error of Non-response: This arises when the respondents do not offer the required information. Error of Misinterpretation: This arises when the respondent fails to interpret the questions in the questionnaire.

Error of Calculation or Arithmetical Error: It occurs in the course of addition, subtraction or multiplication of data.

Error of Sampling Bias: It occurs when, for some reason or the other, a part of target population, cannot be included in the choice of a sample.

Larger the field of investigation or larger the population size, greater is the possibility of errors related to the collection of data, or data acquisition. It must be noted here that a non-sampling error is more serious than a sampling error. Because a sampling error can be minimised by opting for a larger sample size. No such possibility exists in case of non-sampling errors.

Multiple Choice Questions

Select the correct alternative:

1. Census method is suitable for that investigation in which:

(a) the size of population is large

(b) high degree of accuracy is not required

(c) there are widely diverse items

(d) intensive examination of diverse items is not required

2. Which of the following methods is used for the estimation of population in a country?

(a) Census method

(b) Sampling method

(c) Both fa) and (b)

(d) None of these

3. Reliability of sampling data depends on:

(a) size of sample

(b) method of sampling

(c) training of enumerators

(d) all of these

4. For drawing lottery ________________ sampling is used.

(a) random

(b) purposive

(c) stratified

(d) quota

5. Personal bias is possible under:

(a) random sampling

(b) purposive sampling

(c) stratified sampling

(d) quota sampling

6. If the investigator wants to select a sample on the basis of diverse characteristics of the population, which method should he use?

(a) Convenience sampling method

(b) Quota sampling method

(c) Stratified sampling method

(d) Both (b) and (c)

7. Which of the following factor(s) are considered when comparison between sampling and census method is made?

(a) Area of survey

(b) Accuracy of data

(c) Cost of collection

(d) All of these

8. Under random sampling, each item of the universe has __ chance of being selected.

(a) equal

(b) unequal

(c) zero

(d) none of these

Answers

Multiple Choice Questions

1. (c)	**2. (a)**	**3. (d)**	**4. (a)**
5. (b)	**6. (d)**	**7. (d)**	**8. (a)**

CHAPTER 4

ORGANISATION OF DATA

When an investigator collects data for an investigation, these are just raw data. Raw data are not capable of offering any meaningful conclusion. This is just like a lump of clay without any specific shape or identity. Data are to be organised before these are presented for final observations or conclusions. "Organisation of the data refers to the arrangement of figures in such a form that comparison of the mass of similar data may he facilitated and further analysis mas be possible.

An important method of organisation of data is to distribute these into different classes on the basis of their characteristics. This process is called classification of data. It involves conversion of raw data into statistical series in a manner such that some meaningful conclusions can be drawn out of them. The present chapter deals with classification of data, focusing on the conversion of raw data into various types of statistical series.

1. WHAT IS CLASSIFICATION?

There are 200 families in your locality. You have collected data regarding their income, expenditure, education, religion, size of family, etc. But this data will be of little use unless you know how many families are educated and how many are uneducated. How many families earn an income exceeding Rs. 5,000 per month and how many earn Rs. 500 or less per month.

In other words, in order to make the raw data meaningful, these must be classified on the basis of their different characteristics, such as educated families and uneducated families, rich families and poor families, etc.

Each such division of data is called Class and the process by which data is divided into different ‘classes’ on the basis of their similarity or diversity is called “Classification”. Thus, classification is the grouping of related facts into different classes. In the words of Conner, “Classification is the process of arranging things (either actually or nationally) in groups or classes according to their resemblances and affinities, and gives expression to the unity of attributes that may exist amongst a diversity of individuals This definition suggests two important features of classification:

What is the Principal Objective of Classification of Data?

It is to capture and distinctively present the diverse characteristics of data.

(i) Data are divided into different groups. For example, on the basis of education, persons may be classified as **educated** and **uneducated.**

(ii) Data are grouped or classified on the basis of their class similarities. All similar units are put in one class and as the similarity changes, class also changes.

Objectives of Classification

Main objectives of Classification are as under:

(1) Brief and Simple: Main objective of classification is to present data in a form that appears to be brief and simple.

(2) **Utility:** Classification enhances utility of the data as it brings out similarity within the diverse set. of data.

(3) **Distinctiveness:** Classification renders obvious differences among the data more distinctly.

(4) **Comparability:** It makes data comparable and estimative.

(5) **Scientific Arrangement:** Classification facilitates arrangement of data in a scientific manner which increases their reliability.

(6) Attractive and Effective: Classification makes data more attractive and effective.

Characteristics of a Good Classification

(1) Comprehensiveness: Classification of the raw data should be so comprehensive that each and every item of the data gets into some group or class. No item should be left out.

(2) Clarity: Classification of the raw data into classes should be absolutely clear and simple. That is, there should be no confusion about the placement of any item in a group.

(3) **Homogeneity:** All items in a group or class must be homogeneous or similar to each other.

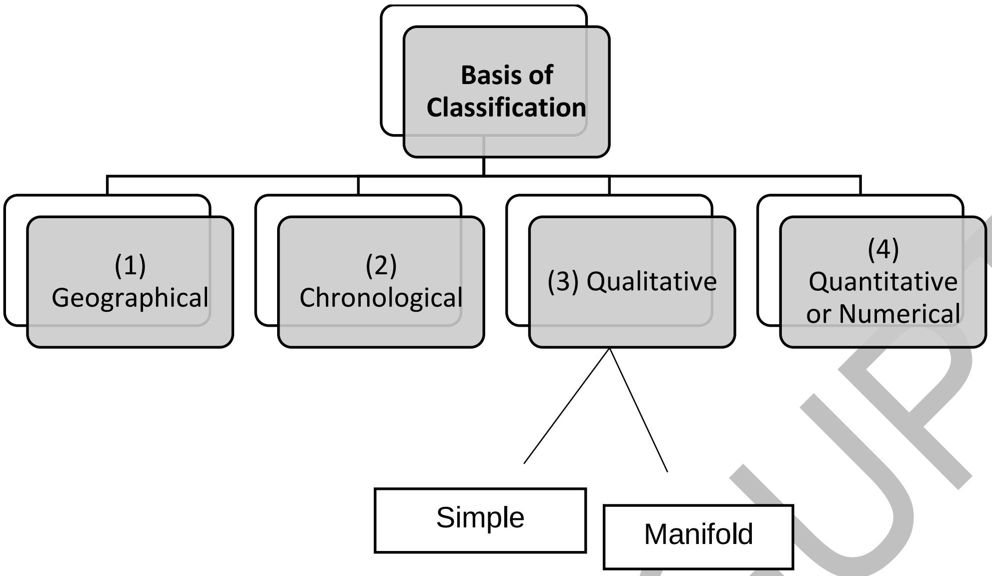
(4) **Suitability:** The composition of the classes must suit the objective of enquiry. For example, in order to determine the income and expenditure of the students in a school, their classification on the basis of weight or marital status would make no sense. The data must be classified on the basis of different levels of income and expenditure.

(5) **Stability:** A particular kind of investigation should be based on the same set of classification. This base should not change with each investigation.

(6) Elastic: Classification should be elastic. There should be a scope for change in the classification, depending on the change in purpose or objective of the study.

Basis of Classification

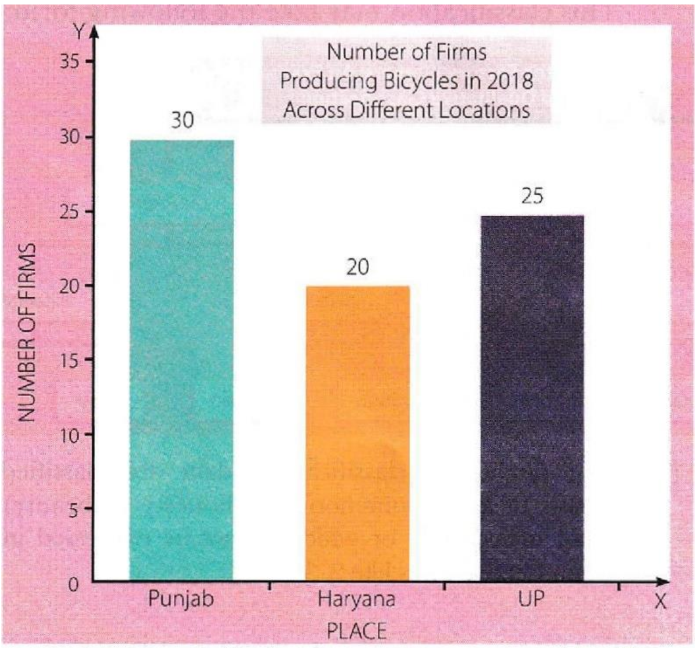
There may be different basis of classifying a statistical information as shown in chart below.



(1) Geographical (or Spatial) Classification: This classification of data is based on the geographical or locational differences of the data. To illustrate, data relating to the number of firms producing bicycles in India would be classified as under:

Table 1. Number of Firms Producing Bicycles in 2018 across Different Locations

Place	Number of Firms
Punjab	30
Haryana	20
UP	25



(2) Chronological Classification: When data are classified on the basis of time, it is known as chronological classification. This is illustrated in the following fable 2.

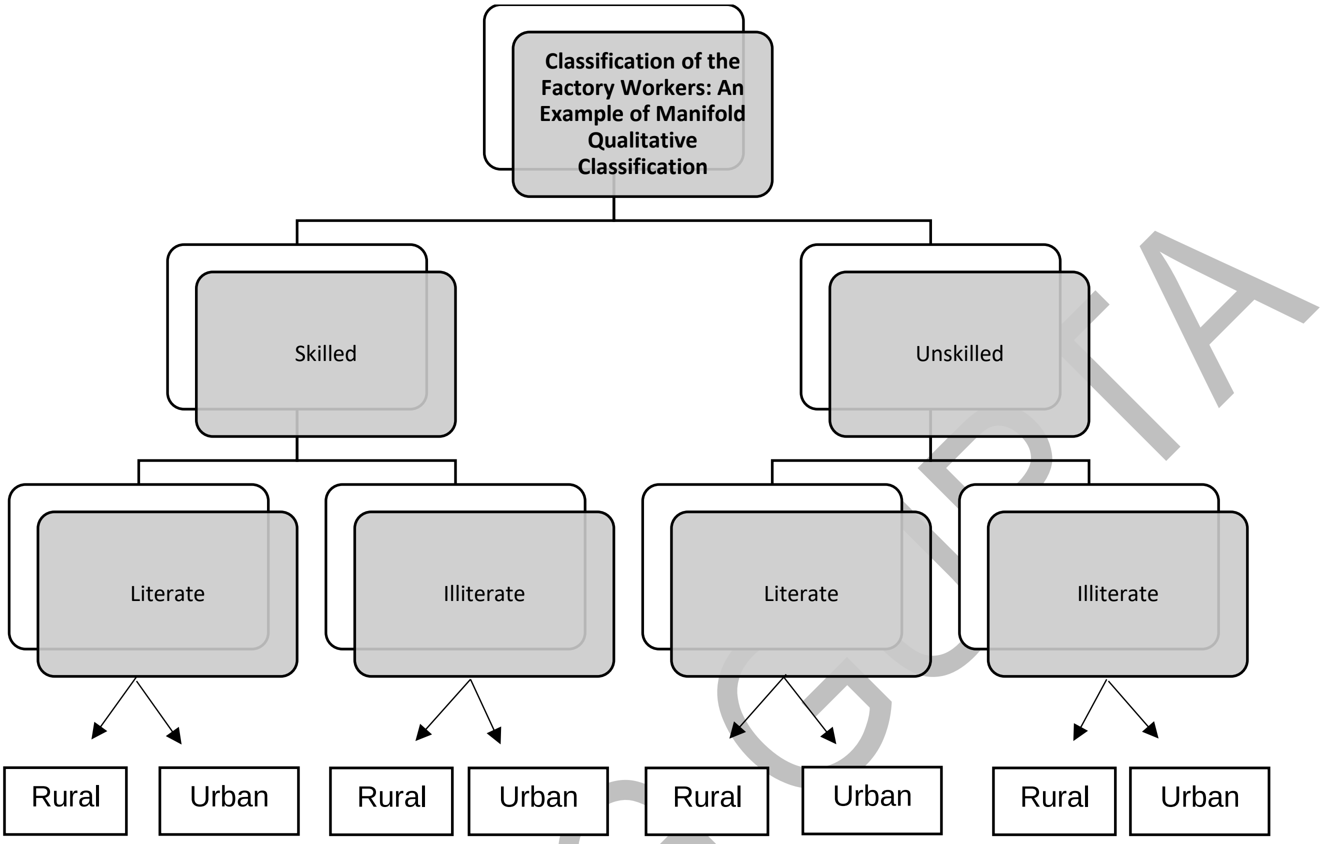
Table 2 Sales of a Firm (2016-2018)

Year	Sales (Rs.)
2016	80 lakh
2017	90 lakh
2018	95 lakh

(3) Qualitative Classification: This classification is according to Qualities or Attributes of the data. For example, data may be classified on the basis of occupation, religion, level of intelligence of the population. This classification may be of two types:

(i) Simple Classification: It is called classification according to dichotomy. This is because data are divided on the basis of existence or absence of a quality. Male-female, healthy-unhealthy, educated-uneducated, are examples of dichotomy.

(ii) Manifold Classification: When classification according to quality of data involves more than one characteristic, it is called manifold classification or multiple classification. As a result of it, there may be more than two classes. To illustrate, factory workers may be classified as ‘skilled’ and ‘unskilled’. These may be further classified as literate or illiterate and still further as rural or urban. This classification may take the following form:



[Note: In qualitative classification, data are classified on the basis of a phenomenon (like honesty or beauty) which is not measurable or which cannot be expressed in terms of quantitative units like 2, 3 or 4.]

(4) Quantitative or Numerical Classification: Classification is done on the basis of numerical values of the facts. A number of classes are framed keeping in view the lowest and highest value as well as the range of values in the data. Each class of a set of data refers to a phenomenon like ‘wages’ or ‘profits’ in the automobile industry which can be expressed in figures like Indian rupees. Table 3 below is an illustration of quantitative classification:

Table 3. Annual Profit of Small-Scale Firms in the State of UP: Hypothetical Data, just for an illustration of Quantitative Classification

Annual Profit (Rs.)	Number of Firms
0-1,00,000	5
1,00,000-2,00,000	150
2,00,000-3,00,000	1500
3,00,000-4,00,000	800

4,00,000-5,00,000	400
Above 5,00,000	200

In the above classification, profit is the phenomenon under study. It is a quantifiable phenomenon. Hence, it is called quantitative classification of data.

It is important to note that the phenomenon under study (like profit in the above illustration) assumes different values over time or across different regions. When a phenomenon assumes different values it is called ‘variable’ in Statistics.

Accordingly, **quantitative classification is also called ‘classification by variables.’**

2. CONCEPT OF VARIABLE

A characteristic or a phenomenon which is capable of being measured and changes its value overtime is called a variable. Thus, a variable refers to that quantity which is subject to change and which can be measured by some unit. If we measure the weight of students of Class XI, then the weight of the students will be called variable. A variable may be either discrete or continuous.

Principal difference between Discrete Variable and Continuous Variable

It is that while discrete variable assumes values in complete numbers like 2, 4, S and 8, continuous variables assume values in fractions like 2.4, 4.6 and 6.8 or values in some range like 2-4, 4-6 etc.

(1) Discrete Variable: Discrete variables are those variables that increase in jumps or in complete numbers. For example, the number of students in Class XI could be 1, 2, 3, 10, 11, 15 or 20 etc. but cannot be $1\frac{1}{4}, 1\frac{1}{2}, 1\frac{3}{4}$ etc. In other words, discrete variables are expressed in terms of complete numbers, or one may simply say that values of these variables are in complete numbers as 1,2, 3 and not continuous as between 1 to $1\frac{1}{4}$ or a $1\frac{3}{4}$ and so on.

(2) Continuous Variable: Variables that assume a range of values or increase not in jumps but continuously or in fractions are called continuous variables. For example, height of the boys in a school is expressed as 5'1", 5'2”, 5'3", and so on.

In short, while the values of discrete variables are in complete numbers (f 2, 3, etc.), values of continuous variables are in fractions (5'4", 5'2", etc.) or are in any range such as JO-15, 15-20, etc.

Difference between Variable and Attribute

Ordinarily, anything that varies/changes over time is taken as a variable. But not in Statistics. Colour of your hair may change over time. Is it a variable? No, not at all. Why? Because this change cannot be numerically expressed. In Statistics, only that change of an object is taken as a variable which can be numerically expressed For example, average height of the students of Class X in the year 2005 is found to be (say) 5'6" compared to 55" in the previous year. Qualitative change, like a change in IQ level of the students of Class X is called attribute'. A change in the attributes is only qualitatively expressed as good, excellent and outstanding. Qualitative changes can at best be ranked as 1, 2, 3 where 1 stands for outstanding, 2 stands for excellent and 3 stands for good.

3. RAW DATA

A mass of data in its crude form is called raw data. It is an unorganised mass of the various items. These are yet to be organised by the investigator.

To illustrate, marks obtained by 30 students of class XI in Statistics, may be expressed as in Table 4.

Table 4 Marks obtained by the Students of Class XI in Statistics

30	20	40	20	15	20
25	10	20	15	25	20
15	45	10	30	20	25
30	20	30	20	15	35
25	10	25	15	35	10

Data presented in this table are raw data. These are not homogeneous data or the data classified into different groups or classes with similarities. No meaningful conclusion is possible from this data. Only that data are useful to Statistics which are homogeneous. An item of the data, like the price of a commodity, income of a farmer, etc. is called observation, or value, or measure, or item, or magnitude, etc. To draw any conclusion from these data, an investigator has to first organise them. To do so, an investigator has to classify the same in the form of series.

Series: Raw data are classified in the form of series. Series refer to those data which are presented in some order and sequence. Arranging of data in different classes according to a given order is called series. Thus, if the marks obtained by the students of Class XI are arranged according to their roll numbers in the ascending or descending order, the data so arranged would be known as statistical series. According to **Horace Secrist,** “4 series as used statistically may be defined as things or attributes of things arranged according to some logical order. ”

Univariate, Bivariate and Multivariate

■ 'Uni' means one, 'bi' means two.' multi means many. Accordingly, **univariate** refers to a series of statistical data with one variable only, like the data on income of the households of a particular region.

■ **Bivariate** refers to a series of statistical data with two variables like the data on income as well as expenditure of the households of a particular region.

- **Multivariate** refers to a series of statistical data with many (and more than two) variables, like the data on age, sex, education, income and expenditure of the households of a particular region.

4. CONVERSION OF RAW DATA INTO STATISTICAL SERIES

Classification of data implies conversion of raw data into statistical series. As stated earlier, classes are formed keeping in mind the nature of variables involved In the study and the range of values that the variables tend to assume.

Types of Statistical Series

Broadly, statistical series are of two types:

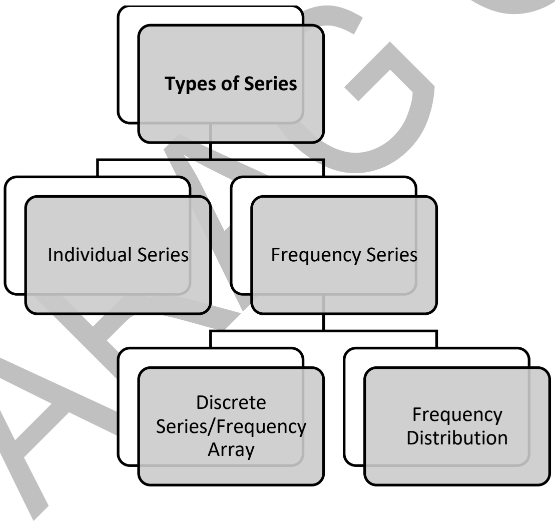
(1) Individual Series or Series without Frequencies, and

(2) Frequency Series or Series with Frequencies

Frequency series are further divided as:

(i) Discrete Series or Frequency Array, and

(ii) Frequency Distribution or Series with Class-Intervals.



(1) Individual Series

Individual series are those series in which the items are listed singly. For example, if the marks obtained by 30 students of Class **XI** are listed singly, the series would be called **Individual Series.** In these series there is no class of the items and also there is no frequency of the items. These series may be presented in two ways:

(i) According to Serial Numbers: One way of presenting an individual series is that all the items are arranged in a serial order. Thus, marks obtained by the students may be arranged in order of their roll numbers. Data on the monthly expenses of the hostel students may be arranged in order of their room numbers. Data given in Table 4 on the marks obtained by 30 students are presented in Table 5 in order of their roll numbers.

Table 5. Marks Obtained by the Students in Statistics

Roll Number	Marks	Roll Number	Marks	Roll Number	Marks	Roll Number	Marks
1	30	9	20	17	15	25	35
2	25	10	10	18	30	26	20
3	15	11	40	19	20	27	20
4	30	12	20	20	15	28	25
5	25	13	10	21	15	29	35
6	20	14	30	22	25	30	10
7	10	15	25	23	20		
8	45	16	20	24	15		

(ii) Ascending or Descending Order of Data: The other way of presenting an individual series is a simple ascending or descending order. In the ascending order, the smallest value is placed first, while in the descending order the highest value is placed first. Tables 6 and 7 show the arrangement of data in the ascending and descending orders respectively.

Table 6. Data Arranged in Ascending Order

10	15	20	25	30
10	15	20	25	30
10	15	20	25	55
10	20	20	25	35
15	20	20	30	40

15	20	25	30	45

Table 7. Data Arranged in Descending Order

45	30	25	20	15
40	30	20	20	15
35	25	20	20	10
35	25	20	15	10
30	25	20	15	10
30	25	20	15	10

Organisation of data in the form of individual series is a very simple form of presentation of data. But this method is not of much use when the number of items is very large.

(2) Frequency Series

Frequency series or series with frequencies may be of two types:

(i) **Discrete Series or Frequency Array**, and

(ii) **Frequency Distribution.**

Before we discuss these two types of series, let us understand the meaning of the following terms:

(a) **Frequency:** Frequency is the number of times an item occurs (or repeats itself) in the series. In other words, the number of times an item repeats itself in the population, is called the frequency of that item. For example, in Table 4, 10 has occurred 4 times. This means 4 students have secured 10 marks; or the frequency of 10 is 4.

(b) **Class Frequency:** The number of times an item repeats itself corresponding to a range of value (or class interval) is called class frequency. For example, if there are 4 students securing marks between 10-15, then 4 is the frequency corresponding to the class interval 10-15. Thus, 4 will be called class frequency.

(c) **Tally Bars:** Every time an item occurs, a tally bar, (|) is marked against that item. Corresponding to a particular class interval, each tally bar signifies ‘one’ occurrence of that item. Two tally bars would mean that the concerned item has occurred twice in the series. After every four tallies the fifth tally will cross out all the previous four tallies. Thus, making a group of five, i.e., ["HJ- This method of marking and counting is known as Four and Cross Method. To illustrate, Table 5 shows that 4 students obtained 10 marks, 5 students obtained 15 marks, 8 students obtained 20 marks, 5 students obtained 25 marks, 4 students obtained 30 marks, 2 students obtained 35 marks, 1 student obtained 40 marks and 1 student obtained 45 marks. All these frequencies have been presented in Table 8, using Four and Cross Method.

Table 8. Four and Cross Method of converting Raw Data into Frequency Series

(data as in Table 5 or 6)

Marks	Tally Bars	Frequency
10	||||	4
15	~~||||~~	5
20	~~||||~~ |||	8
25	~~||||~~	5
30	||||	4
35	||	2
40	|	1
45	|	1

In this table, to express 8 tally bars, first of all four tally bars (||||) are marked, fifth tally bar has been marked across the four ||||. The sign ~~||||~~ signifies that an item occurs five times in the series. Likewise, three tally bars are further marked ||||

to make it equal to eight, i.e., ~~||||~~ + | | | =8. Thus, 1 = |; 2 = ||; 3 = |||; 4 = ||||; 5 = ~~||||~~; 6 = ~~||||~~ |; 7 = ~~||||~~ ||; and so on.

(i) Discrete Series or Frequency Array

A discrete series or frequency array is that series in which data are presented, in a way that exact measurements of items are clearly shown. In such series there are no class intervals, and a particular item in the series is numbered rather than measured with some range.

(ii) Frequency Distribution

It is that series in which items cannot be exactly measured. The items assume a range of values and are placed within the range or limits. In other words, data are classified into different classes with a range, the range is called class intervals. Each item in the series is written against a particular class interval by way of a tally bar. The number of times an item occurs is shown as frequency against the class intervals to which that item belongs.

Frequency Distribution

Marks	Tally Bars	Frequency
10-15	𝍷𝍷𝍷𝍷	4
15-20	𝍸	5
20-25	𝍸 𝍷𝍷𝍷	8
25-30	𝍸	5
30-35	𝍷𝍷𝍷𝍷	4
35-40	𝍷𝍷	2
40-45	𝍷	1
45-50	𝍷	1

It is clear from above table that frequency of class interval 10-15 is 4. It means that there are 4 students who have secured marks between 10-15. Likewise, frequency of class interval 20-25 is 8 which means that there are 8 students who have secured marks between 20-25. But, it is not clear that how many students have secured 10 marks in the class interval 10-15 and how* many have secured 11 and 14 marks in the same class interval.

Size of Class

'Size of Class' refers to size of the class interval, or it refers to width of the class. If 'range' (the difference between highest value and lowest value of the series) is (say) 100 and the number of classes is 20, then size of the class will be 100/20 = 5.

. Thus:

Size of the Class (S) = $\frac{\text{Range (r)}}{\text{No.of Classes (n)}}$

Or, $S = \frac{r}{n} = \frac{100}{20} = 5$

considering the above example.

Note; Size of the class must be such that all values belonging to the particular class-interval tend to converge on the mid-value of the class interval. Only then it becomes an ideal class-size. Otherwise, our result would have a high degree of statistical error.

Some Important Terms

Let us understand some important terms before a detailed study of different types of frequency distribution in the next section.

(i) **Class:** A range of values which incorporate a set of items is called a class. For example, 5-10, 10-15 are the classes.

(ii) **Class Limits:** The extreme values of a class are limits. Every class interval has two limits, lower limit and upper limit. Of the class interval 5-10 in the above example, the lower limit is 5 and the upper limit is 70.

(iii) **Magnitude of a Class Interval:** Magnitude of a class interval is the difference between the upper limit and the lower limit of a class. For example, in a class interval 10-15, the magnitude of the class interval would be 15 - 10 - 5. Thus,

Magnitude of a Class Interval

(i) = Upper limit (l_2) - Lower limit (l_1)

The following formula is used to find out Class Interval.

Formula

$$i = l_2 - l_1$$

Where, i - magnitude of a class interval

l_2 = upper limit of the class interval

l_1 = lower limit of the class interval.

(iv) Mid-value: Mid-value is the average value of the upper and lower limits. It is known by adding up the upper limit and lower limit values and dividing the total by 2. Thus,

Mid-value =

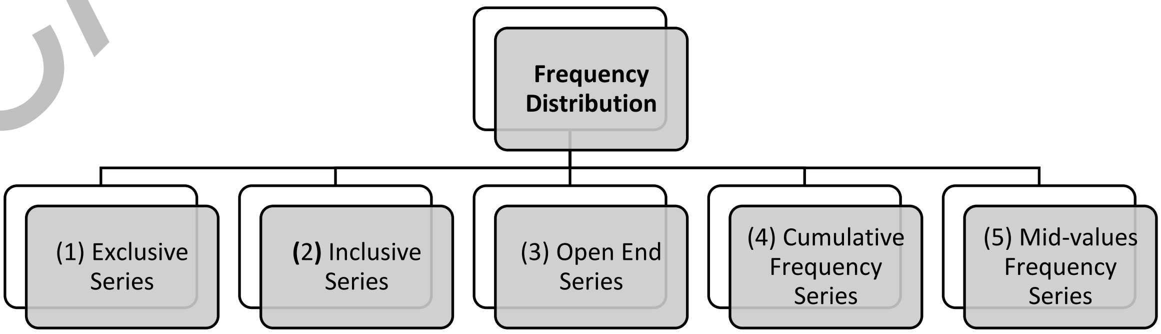
$$\frac{\text{Upper Limit} + \text{Lower Limit}}{2}$$

where, m = mid-value; l_1 = lower limit; l_2 = upper limit.

For example, mid-value of 10-20 class interval $\frac{20+10}{2} = 15$

5. TYPES OF FREQUENCY DISTRIBUTION

Frequency Distribution is of five types:



(1) Exclusive Series

Exclusive series is that series in which every class interval excludes items corresponding to its upper limit. In this series the upper limit of one class interval is the lower limit of the next class interval. It is called exclusive series because frequencies of the upper limit of each class interval is not included in that class.

For example, in a class interval, 10-15, only such items would be included, the value of which is between 10 and 14. Any item of the value of 15 would be included in the next class interval, viz., 15-20.

(2) Inclusive Series

An inclusive series is that series which includes all items upto its upper limit. In such series, the upper limit of class interval does not repeat itself as a lower limit of the next class interval. Thus, there is a gap between the upper limit of a class interval and the lower limit of the next class interval. The gap ranges between 0.1 to 1.0. For example, 10-14, 15-19, 20-24, etc. represents an inclusive series. Thus, all the items ranging between 10-14 are included in that class interval. Likewise, all the items ranging between 15-19 would be included in that class interval.

In short, while in the exclusive series there is an overlapping of the class limits (upper class limit of one class interval being the lower class limit of the next class interval), there is no such overlapping in the inclusive series.

What is the basic difference between Exclusive Series and Inclusive Series?

In case of exclusive series upper limit of one class interval repeats itself as lower limit of the next class interval. While in case of inclusive series, it does not.

Conversion of inclusive Series into Exclusive Series

Inclusive series are used when there is some definite difference between the values of various items in the population. In the above table if a student has obtained, 14.5 or 19.5 marks these can be expressed only if the inclusive series is converted into an exclusive series. Following steps are involved in the conversion of an inclusive series into an exclusive series:

(i) First, we find the difference between the upper limit of class interval and the lower limit of the next class interval.

(ii) Half of that difference is added to the upper limit of a class interval and half is subtracted from the lower limit of the class interval.

Using these two steps, inclusive series of the above table have been converted into an exclusive series as under.

Conversion of the Above Inclusive Series into an Exclusive Series

Marks	Frequency
9.5-14.5	4

14.5-19.5	5
19.5-24.5	8
24.5-29.5	5
29.5-34.5	4

Difference between Exclusive and Inclusive Series

Main differences between exclusive and inclusive series are as under:

(i) In case of exclusive series, the upper limit of one class interval is the lower limit of the next class interval. However, in inclusive series there is generally a difference between the upper limit of one class interval and the lower limit of the other class interval.

(ii) In case of exclusive series, value of the upper limit of a class interval is not included in that class; rather it is included in the lower limit of the next class interval. On the contrary, in the case of inclusive series, value of the upper limit of a class is included in that very class interval.

(iii) Exclusive series is useful whether the value is in complete number or in decimals, but inclusive series is useful only when value is in complete number.

(iv) Counting can be done in all cases under exclusive series. However, to facilitate counting it becomes necessary to convert inclusive series into exclusive series.

(3) Open End Series

In some series, the lower class limit of the first class interval and the tipper limit of the last class interval are missing. Instead, ‘less than or below is specified in place of the lower class limit of the first class interval and ‘more than’ or above is specified in place of the upper class limit of the last class interval. Such series are called ‘Open-end’ series. Thus, an open end series is that series in which lower limit of the first class interval and (or) the upper limit of last class interval is missing. Table below shows such a series.

Open End Series

Marks	Frequency
Below 5	1
5-10	3
10-15	4

15-20	6
20 and above	1

What is an Open-End Series?

It is that series in which

(i) lower limit of the first class interval is missing, or

(ii) upper limit of the last class interval is missing.

In order to determine the limits of the open-end class intervals, the general practice is to give same magnitude to these class intervals as is of the other class intervals in the series. However, this practice is adopted when the known magnitudes of different class intervals in the series are equal to each other. For example, in the above table since the magnitude of the class interval is the same throughout the series, first class interval will be assumed as 0-5 and last as 20-25.

(4) Cumulative Frequency Series

Cumulative frequency series is that series in which the frequencies are continuously added corresponding to each class interval in the series.

Let us proceed with an illustration of converting Simple Frequency Series into a Cumulative Frequency Series. Here is a simple frequency series.

Illustration.

Simple Frequency Series

Marks	Frequency
5-10	3
10-15	8
15-20	9
20-25	4
25-30	4

There are two ways of converting this series into cumulative frequency series. These are:

(i) **Cumulative frequencies may be expressed on the basis of upper limits of the class intervals,** e.g., less than 10, less than 15, less than 20, when the class intervals are 5-10, 10-15 and 15-20.

(ii) **Cumulative frequencies may be expressed on the basis of lower-class limits of the class intervals,** e.g., more than 5, more than 10, more than 15, when the class intervals are 5-10, 10-15 and 15-20.

Thus, when a frequency distribution is to be converted into a cumulative frequency distribution, the cumulative frequencies would correspond to either the lower-class limits or the upper-class limits of the class intervals in a series. Accordingly, the class intervals would get converted into ‘less than’ or ‘more than’ values. Following is an example of how' a frequency distribution is converted into a cumulative frequency distribution.

Cumulative Frequency Series

Method I		**Method II**	
Marks	**Number of Students**	**Marks**	**Number of Students**
Less than 10	0 + 3 = 3	More than 5	28
Less than 15	3 + 8=11	More than 10	28 - 3 = 25
Less than 20	11 + 9 = 20	More than 15	25-8 = 17
Less than 25	20 + 4 = 24	More than 20	17-9= 8
Less than 30	24 + 4 = 28	More than 25	8-4= 4

Conversion of Cumulative Frequency Series into Simple Frequency Series

Cumulative Frequency Series may be converted into Simple Frequency Series. Following illustration explains this process:

Illustration.

Convert the following cumulative frequency series into a simple frequency series.

4 students obtained less than 10 marks

20 students obtained less than 20 marks

40 students obtained less than 30 marks

48 students obtained less than 40 marks

50 students obtained less than 50 marks

Solution:

Conversion of a Cumulative Frequency Series into a Simple Frequency Series

Cumulative Frequency Series		**Simple Frequency Series**	
Marks less than	**Number of Students**	**Marks**	**Number of Students**
10	4	0-10	4
20	20	10-20	20- 4 = 16
30	40	20-30	40 - 20 = 20
40	48	30-40	48-40 = 8
50	50	40-50	50 - 48 = 2

(5) Mid-values Frequency Series

Mid-values frequency series are those series in which we have only mid-values of the class intervals and the corresponding frequencies.

Illustration.

Mid-value	5	15	25	35	45
Frequency	6	5	11	9	8

Such series may be converted into simple frequency series using the following method: (i) First, mutual difference between mid-values (i) is determined; and (ii) Second, the difference so obtained is reduced to half $\left(\frac{1}{2}i\right)$ which when deducted from the mid-value gives lower limit of the class interval and when added to the mid-value gives the corresponding upper limit.

Thus, Lower limit: $l_1 = m - \frac{1}{2}i$

Upper limit: $l_2 = m + \frac{1}{2}i$

where, m = mid-value; i = difference between mid-values; l_1 — lower limit and l_2 = upper limit.

In the above frequency series with raid-values, the mutual difference between mid-values (i) = 15 - 5 = 10. Half of it is 5. Deducting 5 from each mid-value we get lower limits and adding 5 to each mid-value we get the corresponding upper limits.

Following table shows the process of this conversion.

Conversion of a Series with Mid-values into a Series with Class Intervals

Mid-value	Frequency	Classes	Technique
5	6	0-10	$l_1 = 5 - \frac{10}{2} = 0, l_2 = 5 + \frac{10}{2} = 10$
15	5	10-20	$l_1 = 15 - \frac{10}{2} = 10, l_2 = 15 + \frac{10}{2} = 20$
25	11	20-30	$l_1 = 25 - \frac{10}{2} = 20, l_2 = 25 + \frac{10}{2} = 30$
35	9	30-40	$l_1 = 35 - \frac{10}{2} = 30, l_2 = 35 + \frac{10}{2} = 40$
45	8	40-50	$l_1 = 45 - \frac{10}{2} = 40, l_2 = 45 + \frac{10}{2} = 50$

Illustration.

Class mid-values of a frequency distribution of marks in economics of a group of students in Class XI are given as 25, 32, 39, 46, 53 and 60.

Find the size of the class interval and class limits.

Solution:

Size of class interval = Mutual difference between mid-values

= 32 - 25 = 39 - 32 = 46 - 39 - ... = 60-53

= 7

Thus, the size of the class is 7.

Given the size of the class as 7 and mid-values of classes as 25, 32, 39, 46, 53 and 60, we can now obtain the class limits by using the following formulae:

Lower limit: $l_1 = m - \frac{1}{2}i$

Upper limit: $l_2 = m + \frac{1}{2}i$

where, m = mid-values; i — difference between mid-values or the size of class.

Accordingly, the class limits of the first class will be:

$$l_1 = 25 - \frac{1}{2} \times 7 = 25 - 3.5 = 21.5$$

$$l_2 = 25 + \frac{1}{2} \times 7 = 25 + 3.5 = 28.5$$

and so on.

Thus, the various classes with class limits are given as:

Classes	Mid-value	Technique
21.5-28.5	25	$l_1 = 25 - \frac{7}{2} = 21.5, l_2 = 25 + \frac{7}{2} = 28.5$
28.5-35.5	32	$l_1 = 32 - \frac{7}{2} = 28.5, l_2 = 32 + \frac{7}{2} = 35.5$
35.5-42.5	39	$l_1 = 39 - \frac{7}{2} = 35.5, l_2 = 39 + \frac{7}{2} = 42.5$
42.5-49.5	46	$l_1 = 46 - \frac{7}{2} = 42.5, l_2 = 46 + \frac{7}{2} = 49.5$
49.5-56.5	53	$l_1 = 53 - \frac{7}{2} = 49.5, l_2 = 53 + \frac{7}{2} = 56.5$
56.5-63.5	60	$l_1 = 60 - \frac{7}{2} = 56.5, l_2 = 60 + \frac{7}{2} = 63.5$

Multiple Choice Questions

Select the correct alternative:

1. Which of the following is the objective of classification?

(a) Simplification

(b) Briefness

(c) Comparability

(d) All of these

2. Classification of data on the basis of time period is called:

(a) geographical classification

(b) chronological classification

(c) qualitative classification

(d) quantitative classification

3. The characteristic of a fact that can be measured in the form of numbers is called:

(a) frequency

(b) variable

(c) attribute

(d) none of these

4. A series in which every class interval excludes items corresponding to its upper limit is called:

(a) exclusive series

(b) inclusive series

(c) both (a) and (b)

(d) none of these

5. An open-end series is that series in which:

(a) lower limit of the first class interval is missing

(b) upper limit of the last class interval is missing

(c) both (a) and (b)

(d) none of these

6. Formula for finding mid-value is given by:

(a) $I_2 - I_1$

(b) $\frac{l_2 - l_1}{2}$

(c) $I_1 + I_2$

(d) $\frac{l_1 + l_2}{2}$

7. According to tally bar method, which of the following symbols indicate the frequency of five?

(a) IIII

(b) II

(c) III

(d) None of these

8. In a series, the number of times an item occurs is known as:

(a) number

(b) class frequency

(c) frequency

(d) cumulative frequency

9. The difference between upper limit and lower limit of a class is known as:

(a) range

(b) magnitude of a class interval

(c) frequency

(d) class limits

10. Which of the following equations is correct?

(a) s = r + n (b) s = r - n

(c) s = r × n (d) s = $\frac{r}{n}$

[s = Size of the class; r = Range; n = Number of classes]

11. Drinking habit of a person is:

(a) An attribute

(b) A discrete variable

(c) A variable

(d) A continuous variable

12. An attribute is:

(a) A qualitative characteristic

(b) A measurable characteristic

(c) A quantitative characteristic

(d) All these

13. Nationality of a student is:

(a) An attribute

(b) A discrete variable

(c) A continuous variable

(d) Either (a) or (c)

14. For the construction of a grouped frequency distribution, we take:

(a) Class boundaries

(b) Class limits

(c) Both (a) and (b)

(d) None of these

15. Tally marks determines:

(a) Class width

(b) Class boundary

(c) Class limit

(d) Class frequency

16. The number of observations falling within a class is called:

(a) Density

(b) Frequency

(c) Both (a) and (b)

(d) None of these

17. A series showing the sets of all distinct values individually with their frequencies is known as:

(a) Grouped frequency distribution

(b) Simple frequency distribution

(c) Cumulative frequency distribution

(d) None of these

18. Annual income of a person is:

(a) A continuous variable

(b) A discrete variable

(c) An attribute

(d) Either (b) or (c)

19. Upper limit of any class is:

(a) Same

(b) Different

(c) Both (a) and (b)

(d) None of these

20. In inclusive class-intervals of a frequency distribution:

(a) Upper limit of each class-interval is included

(b) Lower limit of each class-interval is included

(c) Both (a) and (b)

(d) None of these

21. In exclusive class intervals of a frequency distribution:

(a) Upper limit of each class-interval is excluded

(b) Lower limit of each class-interval is excluded

(c) Both (a) and (b)

(d) None of these

22. For determining the class frequencies, it is necessary that these classes are:

(a) Mutually exclusive

(b) Not mutually exclusive

(c) Independent

(d) None of these

23. The Frequency distribution of a continuous variable is known as:

(a) Grouped frequency distribution

(b) Simple frequency distribution

(c) Either (a) or (b)

(d) Both (a) and (b)

24. In an ordered series, the data are:

(a) In descending order

(b) In ascending order

(c) Either (a) or (b)

(d) None of these

25. The value exactly at the middle of a class-interval is called:

(a) Class mark

(b) Mid-value

(c) Both (a) and (b)

(d) None of these

26. The lower-class boundary is:

(a) An upper limit to Lower Class Limit

(b) A Lower limit to Lower Class Limit

(c) Both (a) and (b)

(d) None of these

27. The upper limit of class-intervals is considered for calculating:

(a) ‘Less than’ cumulative frequency

(b) ‘More than’ cumulative frequency

(c) Relative frequency

(d) None of these

28. In an individual series, each variate value has:

(a) Same frequency

(b) Frequency one

(c) Varied frequency

(d) Frequency two

29. A grouped frequency distribution with uncertain first or last class is known as:

(a) Exclusive class distribution

(b) Inclusive class distribution

(c) Open end distribution

(d) Discrete frequency distribution

30. The following data relate to the marks of a group of students:

Marks	No. of students
Below 10	15
Below 20	38
Below 30	65
Below 40	84
Below 50	100

How many students get marks more than 30?

(a) 65	(b) 50
(c) 35	(d) 43

31. Class-interval is measured as:

(a) Half of the sum of lower and upper limit

(b) The sum of the upper and lower limit

(c) Half of difference between upper and lower limit

(d) The difference between upper and lower limit

32. The data given as 5, 7,12,17, 79, 84, 91 will be called as:

(a) A continuous series

(b) A discrete series

(c) An individual series

(d) Time series

33. Most extreme values which are never included in a class-interval are called:

(a) Class-interval

(b) Class limits

(c) Cass boundaries

(d) None of these

34. The class marks of a distribution are 26, 31,36, 41,46 and 51. Then the first-class interval is:

(a) 23.5-28.5

(b) 23-28

(c) 22.5-27.5

(d) None of these

35. Mutually exclusive classification:

(a) Excludes the upper-class limit but includes the lower-class limit

(b) Excludes both the class limits

(c) Includes the upper-class limit but excludes the upper-class limit

(d) Either (b) or (c)

36. In the construction of a frequency distribution, it is generally preferable to have classes of

(a) Equal width

(b) Unequal width

(c) Maximum width

(d) None of these

37. Classes with zero frequencies are called:

(a) Class

(b) Empty class

(c) Nil class

(d) None of these

38. Frequency of a variable is always:

(a) A fraction

(b) In percentage

(c) An integer

(d) None of these

39. Find the number of observations between 250 and 300 from the following data:

Value	No. of observations

More than 200	56
More than 250	38
More than 300	15
More than 350	0

(a) 56 (b) 23

(c) 15 (d) 8

40. A series showing the sets of all values in classes with their corresponding frequencies is known as:

(a) Grouped frequency distribution

(b) Cumulative frequency distribution

(c) Simple frequency distribution

(d) None of the above

41. For the mid-values given: 25, 34, 43, 53, 61,70, the first class of the distribution is:

(a) 25-34

(b) 24.5-34.5

(C) 20-30

(d) 20.5-29.5

42. Why is it true that classes in frequency distributions are all inclusive`

(a) No data point falls into more than one class

(b) There are always more classes than data point

(c) All data fit into one class or another

(d) All of these

43. Mutually exclusive classification is usually meant for:

(a) An attribute

(b) A continuous variable

(c) A discrete variable

(d) Any of these

Answers

Multiple Choice Questions

1. (d)	**2. (b)**	**3. (b)**	**4. (a)**	**5. (c)**
6. (d)	**7. (d)**	**8. (c)**	**9. (b)**	**10. (d)**
11. (a)	**12. (a)**	**13. (a)**	**14. (b)**	**15. (a)**
16. (b)	**17. (b)**	**18. (b)**	**19. (b)**	**20. (c)**
21. (a)	**22. (a)**	**23. (a)**	**24. (c)**	**25. (c)**
26. (b)	**27. (a)**	**28. (b)**	**29. (c)**	**30. (c)**
31. (d)	**32. (c)**	**33. (c)**	**34. (a)**	**35. (b)**
36. (a)	**37. (b)**	**38. (c)**	**39. (b)**	**40. (a)**
41. (d)	**42. (c)**	**43. (b)**		

CHAPTER 5

PRESENTATION OF DATA— TEXTUAL AND TABULAR PRESENTATION

The presentation of data means exhibition of the data in such a dear and attractive manner that these are easily understood and analysed. There are many forms of presentation of data of which the following three are well known: (i) Textual or Descriptive Presentation, (ii) Tabular Presentation, and (iii) Diagrammatic Presentation. The present chapter focuses on Textual and Tabular Presentation of data. Diagrammatic Presentation of data is discussed in the next chapter.

1. TEXTUAL PRESENTATION

In textual presentation, data are a part of the text of study or a part of the description of the subject matter of study. Such a presentation is also called descriptive presentation of data. This is the most common form of data presentation when the quantity of data is not very large. Here are some examples:

Example 1

In a strike call given by the trade unions of shoe making industry in the city of Delhi, 50% of the workers reported for the duty, and only 2 out of the 20 industries in the city were totally closed.

Example 2

Surveys conducted by a Non-government Organisation reveal that, in the state of Punjab, area under pulses has tended to shrink by 40% while the area under rice and wheat has tended to expand by 20%, between the years 2001-2011.

Suitability

Textual presentation of data is most suitable when the quantum of data is not very large. A small volume of data presented as a part of the subject matter of study becomes a useful supportive evidence to the text. Thus, rather than saying that price of gold is sky-rocketing, a statement like price of gold has risen by 50% during the financial year 2017-18 is much more meaningful and precise. One need not support the text with voluminous data in the form of tables or diagram when the textual matter itself is very small and includes only a few observations. Indeed, textual presentation of data is an integral component of a small quantitative description of a phenomenon. It gives an emphasis of statistical truth to the otherwise qualitative observations.

Drawbacks

A serious drawback of die textual presentation of data is that one has to go through the entire text before quantitative facts about a phenomenon become evident. A picture or a set of bars showing increase in the price of gold during a specified period is certainly quite informative even on a casual glance of the reader. Textual presentation of data, on the other hand, does not offer anything to the reader at a mere glance of the text matter. The reader must read and comprehend (he entire text. When the subject under study is vast and involves comparison across different areas/countries, textual presentation of data would only add to discomfort of the reader.

2. TABULAR PRESENTATION

In the words of Neiswanger, "A statistical table is a systematic organisation of data in columns and rows" Vertical dissections of table (||) are known as columns and horizontal dissections (=) are known as rows.

Tabulation is the process of presenting data in the form of a table. According to Prof. L.R. Connor, 'tabulation involves the orderly and systematic presentation of numerical data in a form designed to elucidate the problem under consideration. "

In the words of Prof. M.M. Blair, "Tabulation in its broadest sense is an orderly arrangement of data in columns and rows."

Components of a Table

Following are the principal components of a table:

(1) Table Number: First of all, a table must be numbered. Different tables must have different numbers, e.g., 1, 2, 3, etc. These numbers must be in the same order as the tables. Numbers facilitate location of the tables.

(2) Title: A table must have a title. Title must be written in bold letters. It should attract the attention of the readers. The title must be simple, clear and short. A good title must reveal: (i) the problem under consideration, (ii) the time period of the study, (iii) the place of study, and (iv) the nature of classification of data. A good title is short but complete in all respects.

(3) **Head Note:** If the title of the table does not give complete information, it is supplemented with a head note. Head note completes the information in the title of the table. Thus, units of the data are generally expressed in the form of lakhs, tonnes, etc. and preferably in brackets as a head-note.

(4) Stubs: Stubs are titles of the rows of a table. These titles indicate information contained in the rows of the table.

(5) Caption: Caption is the title given to the columns of a table. A caption indicates information contained in the columns of the table.

A caption may have sub-heads when information contained in the columns is divided in more than one class. For example, a caption of ‘Students’ may have boys and girls as sub-heads.

(6) Body or Field: Body of a table means sum total of the items in the table. Thus, body is the most important part of a table. It indicates values of the various items in the table. Each item in the body is called ‘cell’.

(7) Footnotes: Footnotes are given for clarification of the reader.

These are generally given when information in the table need to be supplemented. «

(8) Source: When tables are based on secondary data, source of the data is to be given. Source of the data is specified below the footnote. It should give: name of the publication and publisher, year of publication, reference, page number, etc.

Difference between Table and Tabulation

While tabulation refers to the method or process of presenting data in the form of rows and columns, table refers to the actual presentation of data in the form of rows and columns. Table is the consequence (result) of tabulation.

Check [he following format of a table showing its various components:

FORMAT OF TABLE

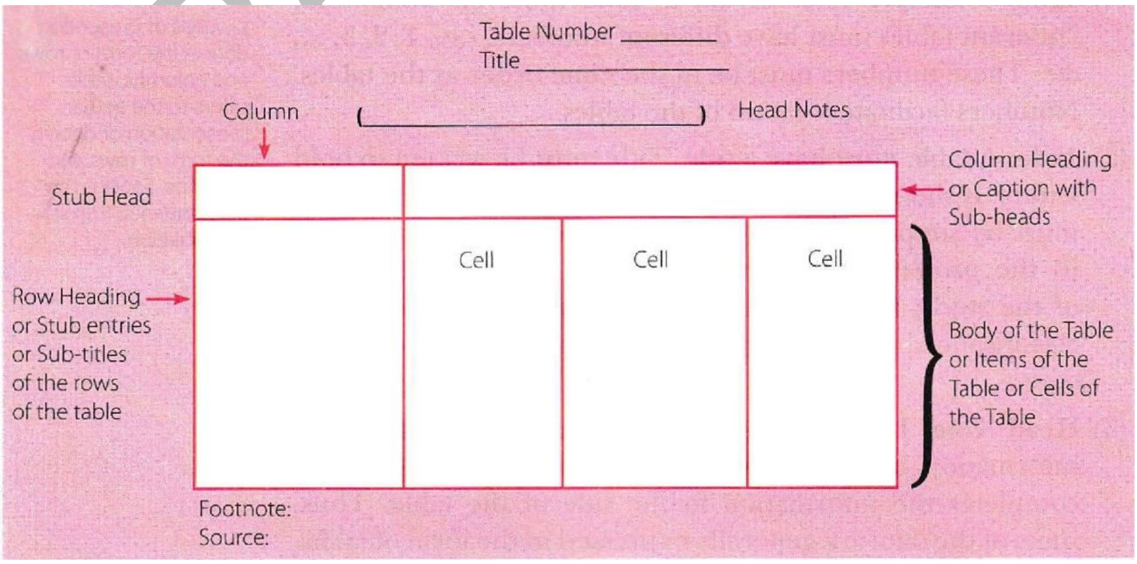


Table Number ________
Title ______________
Column
(______________)
Head Notes
Stub Head
Column Heading or Caption with Sub-heads
Cell
Cell
Cell
Row Heading or Stub entries or Sub-titles of the rows of the table
Body of the Table or Items of the Table or Cells of the Table
Footnote:
Source:

Guidelines for the Construction of a Table or Features of a Good Table

Construction of a table depends upon the objective of study. It also depends upon the wisdom of the statistician. There are no hard and fast rules for the construction of a table. However, some important guidelines should be kept in mind. These guidelines are features of a good table. These are as under:

(1) Compatible Title: Title of a table must be compatible with the objective of the study. The title should be placed at the top centre of the table.

(2) Comparison: It should be kept in mind that items (cells) which are to be compared with each other are placed in columns or rows close to each other. This facilitates comparison.

(3) **Special Emphasis:** Some items in the table may need special emphasis. Such items should be placed in the head rows (top above) or head columns (extreme left). Moreover, such items should be presented in bold figures.

(4) Ideal Size: Table must be of an ideal size. To determine an ideal size of a table, a rough draft or sketch must be drawn. Rough draft will give an idea as to how many rows and columns should be drawn for presentation of the data.

(5) Stubs: If rows are very long, stubs may be given at the right hand side of the table also.

(6) Use of Zero: Zero should be used only to indicate the quantity of a variable. It should not be used to indicate the non-availability of data. If the data are not available, it should be indicated by ‘n.a.’ or (-) hyphen sign.

(7) Headings: Headings should generally be written in the singular form. For example, in the columns indicating goods, the word ‘good’ should be used.

(8) Abbreviations: Use of abbreviations should be avoided in the headings or sub-headings of the table. Short forms of the words such as Govt., m.p. (monetary policy), etc. should not be used. Also such signs as “(ditto)” should not be used in the body of the table.

(9) Footnote: Footnote should be given only if needed. However, if footnote is to be given, it must bear some asterisk mark (*) corresponding to the concerned item.

(10) Units: Units used must be specified above the columns. If figures are very large, units may be noted in the short form as ‘000’ hectare or ‘000’ tonnes.

(11) Total: In the table, sub-totals of the items must be given at the end of each row. Grand total of the items must also be noted.

(12) Percentage and Ratio: Percentage figures should be provided in the table, if possible. This makes the data more informative.

(13) Extent of Approximation: **If** some approximate figures have been used in the table, the extent of approximation must be noted. This may be indicated at the top of the table as a part of head note or at the foot of the table as a footnote.

(14) Source of Data: Source of data must be noted at the foot of the table. It is generally noted next to the footnote.

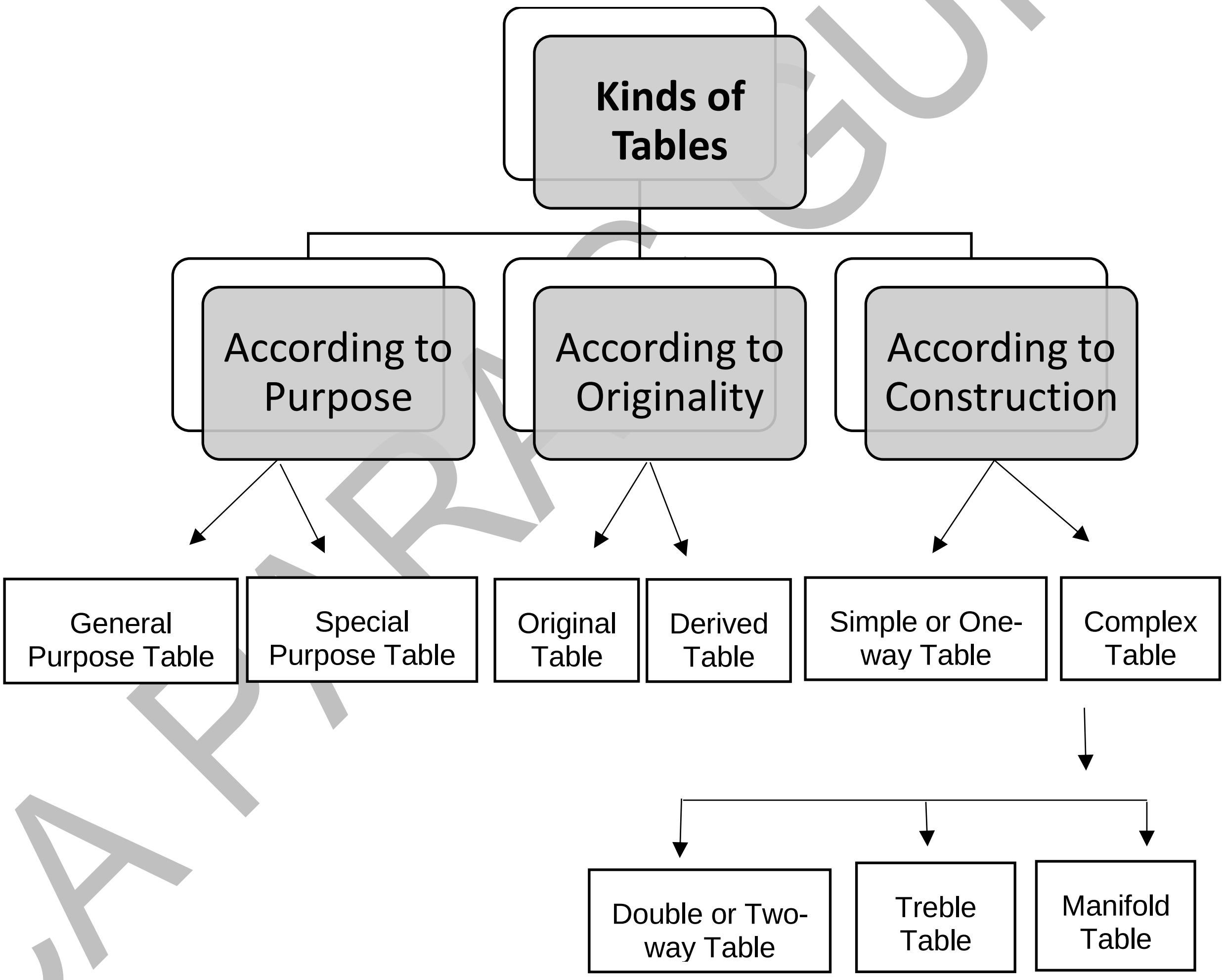
(15) Size of Columns: Size of the columns must be uniform and symmetrical.

(16) Ruling of Columns: Columns may be divided into different sections according to similarities of the data.

(17) Simple, Economical and Attractive: A table must be simple, attractive and economical in space.

Kinds of Tables

There are three basis of classifying tables, viz., (1) purpose of a table, (2) originality of a table, and (3) construction of a table. According to each of these bases, statisticians have classified tables as in the following flow chart:



Let us attempt a brief description of the various kinds of tables:

(1) Tables according to Purpose

According to purpose, there are two kinds of tables:

(i) **General Purpose Table:** General purpose table is that table which is of general use. It does not serve any specific purpose or specific problem under consideration. Such tables are just ‘data bank’ for the use of researchers for their various studies. These tables are generally attached to some official reports, like Census Reports ofIndia. These are also called **Reference Tables.**

(ii) Special Purpose Table: Special purpose table is that table which is prepared with some specific purpose in mind. Generally, these are small tables limited to the problem under consideration. In these tables data are presented in the form of result of the analysis. That is why these tables are also called **summary tables.**

(2) Tables according to Originality

On the basis of originality, tables are of two kinds:

(i) Original Table: An original table is that in which data are presented in the same form and manner in which they are collected.

(ii) Derived Table: A derived table is that in which data are not presented in the form or manner in which these are collected. Instead the data are first converted into ratios or percentage and then presented.

(3) Tables according to Construction

According to construction, tables are of two kinds:

(i) Simple or One-way Table: A simple table is that which shows only one characteristic of the data. Table 2 below is an example of a simple table. It shows number of students in a college:

Number of Students in a College

Class	Number of Students
XI	200
B.A. (I)	100
B.A. (II)	80
B.A. (III)	60
Total	440

(ii) **Complex Table:** A complex table is one which shows more than one characteristic of the data. On the basis of the characteristics shown, these tables may be further classified as:

(a) Double or Two-way Table: A two-way table is that which shows two characteristics of the data. For example, Table 3, showing the number of students in different classes according to their sex, is a two-way table:

Number of Students in a College

(According to Sex and Class)

Class	**Number of Students**		**Total**
	Boys	**Girls**	
XI	160	40	200
B.A. (I)	40	60	100
B.A. (11)	60	20	80
B.A.(III)	50	10	60
Total	310	130	440

(b) Treble Table: A treble table is that which shows three characteristics of the data. For example, Table 4 shows number of students in a college according to class, sex and habitation.

Number of Students in a College

(According to Class, Sex and Habitation)

Class	**Boys**			**Girls**			**Total**		
	Rural	**Urban**	**Total**	**Rural**	**Urban**	**Total**	**Rural**	**Urban**	**Total**
XI	50	110	160	10	30	40	60	140	200
B.A. (I)	10	30	40	15	45	60	25	75	100
B.A. (IT)	15	45	60	5	15	20	20	60	80
B.A. (III)	10	40	50	5	5	10	15	45	60
Total	85	225	310	35	95	130	120	320	440

(c) **Manifold Table:** A manifold table is the one which shows more than three characteristics of the data. Table 5, for example, shows number of students in a college according to their sex, class, habitation and marital status.

Number of Students in a College

(According to their Sex, Class, Habitation and Marital Status)

Class	Boys				Girls				Total
	Rural		Urban		Rural		Urban		
	Married	Unmarried	Married Unmarried		Married Unmarried		Married Unmarried		
XI	5	55	10	90	2	8	5	25	200
B.A. (1)	5	15	15	35	4	4	4	18	100
B.A. (II)	5	10	15	30	2	3	5	10	80
B.A.(III)	5	5	20	20	3	2	2	3	60
Total	20	85	60	175	11	17	16	56	440

Classification of Data and Tabular Presentation

Tabular presentation is based on four-fold classification of data, viz., qualitative, quantitative, temporal, and spatial. Following are the details with suitable illustrations.

(1) Qualitative Classification of Data and Tabular Presentation:

Qualitative classification occurs when data are classified on the basis of qualitative attributes or qualitative characteristics of a phenomenon. **Example:** Data of unemployment may relate to rural-urban areas, skilled and unskilled workers, or male and female job-seekers. Table 6 below is an example of tabular presentation of data when data are classified on the basis of qualitative attributes or qualitative characteristics.

Unemployment in Punjab by Sex and Location (per cent)

Sex	Location	
	Rural	Urban

Male	20	10
Female	30	20
Total	25	15

(This is an imaginary table. In this table, male and female are such characteristics/attributes which are qualitative and cannot be quantified.)

(2) Quantitative Classification of Data and Tabular Presentation:

Quantitative classification occurs when data are classified on the basis of quantitative characteristics of a phenomenon.

Example: Data on marks in Mathematics by the students of Class XII in CBSE examination. Table 7 shows tabular presentation of data when data are classified on the basis of quantitative characteristics.

Marks Obtained by Students of Class XII of XYZ School

Marks	**Number of Students**
20-30	3
30-40	7
40-50	12
50-60	22
60-70	32
70-80	36
80-90	09
90-100	01

Source: Result Sheets

Here, marks are a quantifiable variable and data are classified in terms of different class intervals of marks.

(3) Temporal Classification of Data and Tabular Presentation:

In temporal classification, data are classified according to time, and time becomes the classifying variable.

Example: Sale of Cell phones in different years during the period 2014-2018 in the city of Delhi. Table 8 shows tabular presentation of data on the basis of temporal classification.

Annual Sale of Cell Phones in the City of Delhi (2014-2018)

Year	Sale (Units)
2014	50,000
2015	70,000
2016	90,000
2017	1,00,000
2018	2,00,000

(4) Spatial Classification:

In spatial classification, place/location becomes the classifying variable. It may be a village, a town, a district, a state or a country as a whole.

Example: Number of Indian students studying in different countries of the world during a particular year. Table 9 is an example of tabular presentation based on spatial classification of data.

Indian Students in different Countries of the World (2018)

Country	Number of Students
USA	50,000
UK	15,000
Japan	5,000
Russia	2,000

Australia	7,000

Merits of Tabular Presentation

Following are the principal merits of tabular presentation of data:

(1) Simple and Brief Presentation: Tabular presentation is perhaps the most simplest form of data presentation. Data, therefore, are easily understood. Also, a large volume of statistical data is presented in a very brief form.

(2) Facilitates Comparison: The tabulation facilitates comparison of data by presenting the data in different classes.

(3) **Easy Analysis:** It is very easy to analyse the data from tables. It is by organising the data in the form of table that one finds out their central tendency, dispersion and correlation.

(4) Highlights Characteristics of Data: Tabulation highlights characteristics of data. Accordingly, it becomes easy to remember the statistical facts.

(5) **Economical:** Tabular presentation is a very economical mode of data presentation. It saves time as well as space.

Multiple Choice Questions

Select the correct alternative:

1. The process of presenting data in the form of a table is called:

(a) organisation

(b) classification

(c) presentation

(d) tabulation

2. The principal component of a table is:

(a) table number

(b) title

(c) head note

(d) all of these

3. Which of the following is a basis of classification of a table?

(a) Purpose

(b) Construction

(c) Originality

(d) All of these

4. Which of the following are titles of the rows of a table?

(a) Title

(b)Stub

(c) Caption

(d) None of these

5. Complex table may be classified as:

(a) general purpose and special purpose table

(b) original and derived

(c) double, treble and manifold table

(d) none of these

6. In temporal classification, data are classified on the basis of:

(a) location

(b) time

(c) originality

(d) purpose

7. Table is the consequence of:

(a) classification

(b) organisation

(c) presentation

(d) tabulation

Answers

Multiple Choice Questions

1.(d) **2. (d)** **3. (d)** **4. (b)** **5. (c)**

6. (b) **7. (d)**

CHAPTER 6

DIAGRAMMATIC PRESENTATION OF DATA— BAR DIAGRAMS AND PIE DIAGRAMS

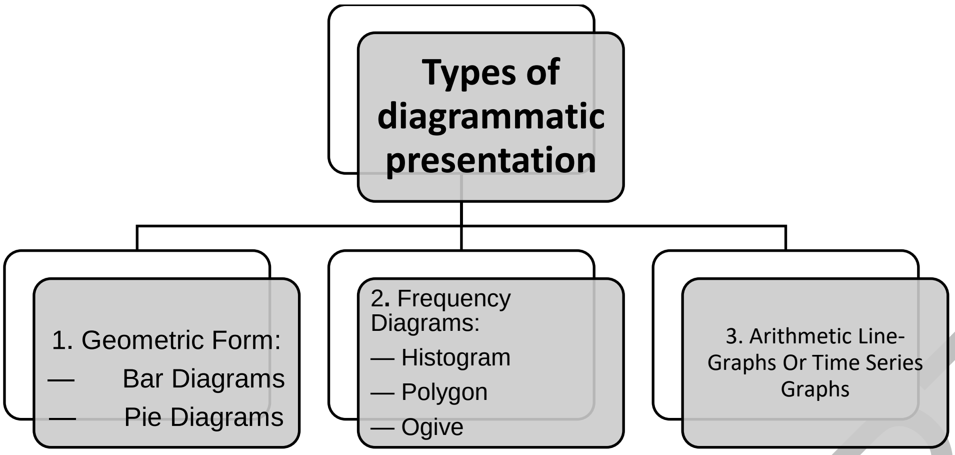
Data may be presented in a simple and attractive manner in the form of diagrams. Diagrammatic presentation is broadly classified as of three types, viz.,

(i) Geometric form including bar diagrams and pie diagrams,

(ii) Frequency diagrams including histograms, polygon and ogive, and

(iii) Arithmetic Line-Graphs or Time Series Graphs.

Following chart shows different types of diagrammatic presentation as specified in the syllabus for the CBSE students of Class XI.



Types of diagrammatic presentation

1. Geometric Form:
— Bar Diagrams
— Pie Diagrams

2. Frequency Diagrams:
— Histogram
— Polygon
— Ogive

3. Arithmetic Line-Graphs Or Time Series Graphs

The present chapter focuses on Geometric form of diagrammatic presentation, including (i) Bar diagrams, and (ii) Pie diagrams.

1. BAR DIAGRAMS

Bar diagrams are those diagrams in which data are presented in the form of bars or rectangles. Bars are also called columns.

Bar and its Features

Bar usually means a rectangle or some rectangular form. It shows some value of the variable. It has the following features:

(i) The length or height of the bars differs according to different values of the variable, that is, length may be more or less but breadth remains the same,

(ii) Bars may be either vertical or horizontal. However, usually these are used in their vertical form.

(iii) Bars are equidistant from each other.

(iv) All bars are based on some common base line.

(v) Unless data assume some specific ordering, these should be presented to form bars of the ascending or descending order.

(vi) To make the bars attractive, these may be shaded with different colours.

Types of Bar Diagrams

(1) Simple Bar Diagrams: Simple bar diagrams are those diagrams which are based on a single set of numerical data. The different items or values are represented by different bars.

Data relating to birth rate, death rate, production level of a commodity, are generally presented in the form of simple bar diagrams. These bar diagrams present only a single set of numerical data. These diagrams are more useful to present time series data, such as, individual value like weight of students; time series like birth-rate according to 2001-2011 census surveys; geographical conditions like per capita income in different states of India, namely, Haryana, Punjab, Himachal Pradesh, etc. The main limitation of simple bar diagrams is that these diagrams can show only a single set of numerical data. Following illustrations explain the construction of simple bar diagram.

Illustration.

The following table gives data on birth rate in India according to census survey of different years. Present the information in the form of a vertical/simple bar diagram.

Year	1951-60	1961-70	1971-80	1981-90	1991-2000	2001-2010
Birth Rate	45	35	30	28	24	20

Solution; **Birth Rate in India**

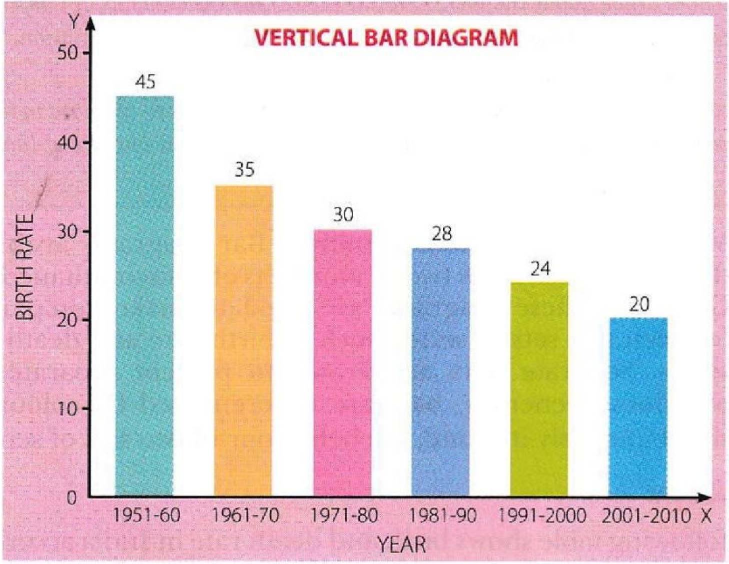


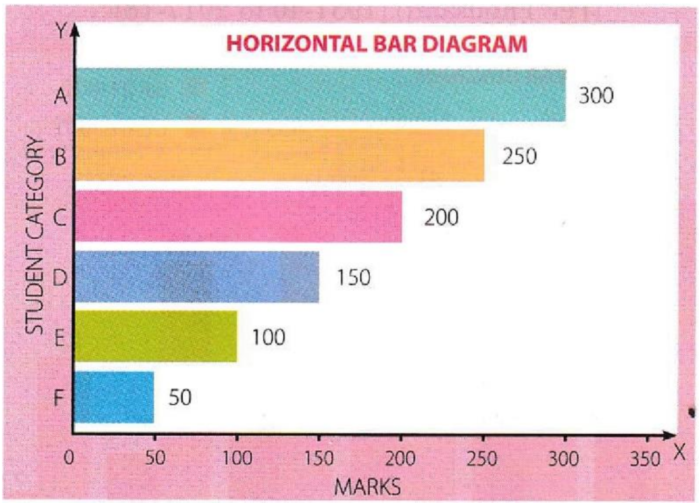
Illustration.

Present the following data in the form of horizontal bar diagram:

Students	A	B	C	D	E	F
Marks	300	250	200	150	100	50

Solution:

Marks obtained by different categories of students:



In the above diagram data are presented in the form of horizontal bars. All the bars are of equal breadth. These are also equidistant from each other.

Illustration.

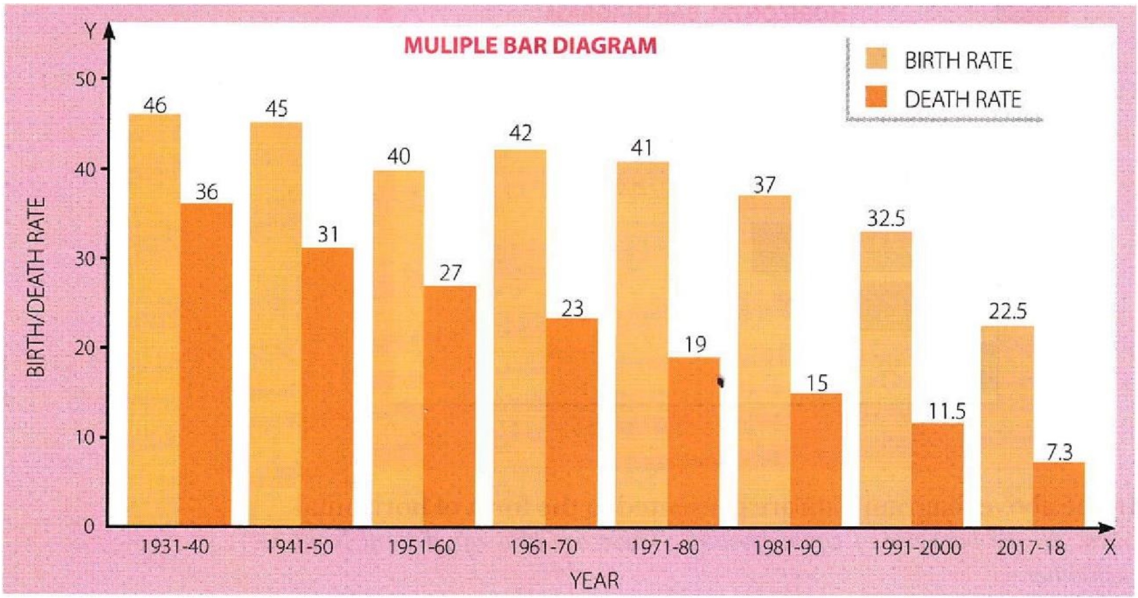
The following table shows birth and death rate in India according to the Census Reports between 1931-40 to 2016-17 (hypothetical figures). Present the data in the form of a multiple bar diagram.

Year	1931-40	1941-50	1951-60	1961-70	1971-80	1981-90	1991-2000	2017-18
Birth Rate	46	45	40	42	41	37	32.5	22.5
Death Rate	36	31	27	23	19	15	11.5	7.3

Solution:

Birth Rate and Death Rate in India

(Per Thousand) (1931-40 to 2017-18)



(3) Sub-divided Bar Diagrams or Differential Bar Diagrams: Sub-divided bar diagrams are those diagrams which simultaneously present, total values as well as part values of a set of data. Different parts of the bars are shaded with colours. However, different parts of a bar must be shown in the same order for all bars of a diagram.

Illustration.

The following table shows Production of Electricity from different Sources in India during 2014-15 to 2017-18 (hypothetical data). Present the data in a sub-divided bar diagram.

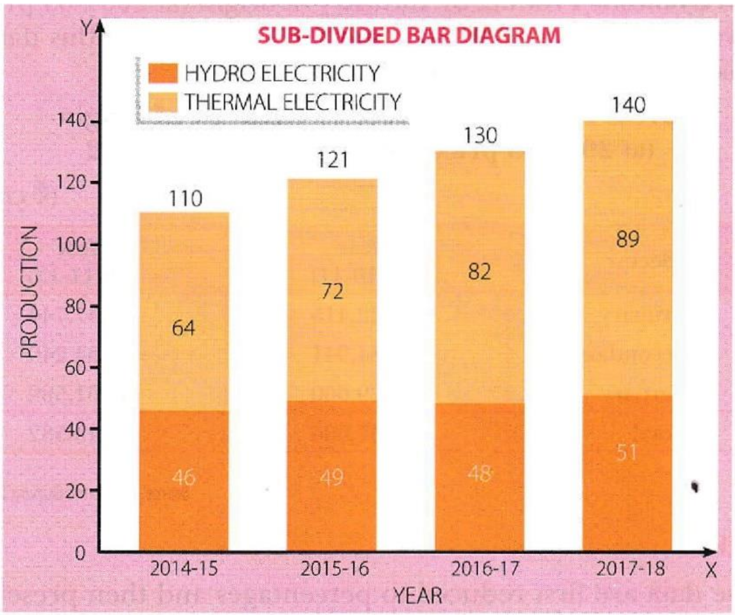
(000’ Million KWTs)

Year	**Hydro Electricity**	**Thermal Electricity**	**Total Production**
2014-15	**46**	**64**	**110**
2015-16	**49**	**72**	**121**
2016-17	**48**	**82**	**130**
2017-18	**51**	**89**	**140**

Solution:

Hydro/Thermal Electricity Production in India

2014-15 to 2017-18 (000’ Million KWTs)



(4) Percentage Bar Diagrams: Percentage bar diagrams are those diagrams which show simultaneously, different parts of the values of a set of data in terms of percentages. Total value, indicated by total length of a bar is assumed to be 100. Each part thereof is shown as a part of 100. These parts may be shaded with different colours in order to highlight the difference. In fact, these diagrams are a different form of sub-divided bar diagrams. The percentage bar diagrams are used, generally when the values are of high magnitude. In such cases, the use of subdivided bar diagrams may not be appropriate. Because, when the values are of high magnitude, some parts of the bar may look very small compared to the others. Wide differences in parts of the bars makes comparison difficult. Percentage distribution of the bars in such cases should certainly be a better mode of presenting the data. Example given below illustrates percentage bar diagram.

Illustration.

Gross Domestic Product by Industry of origin (at 2004-05 prices) is given for the years 2010-11 and 2011-12. Present this data in terms of percentage bar diagram.

Gross Domestic Product by Industry of Origin (at 2004-05 prices)

in 2010-11 and 2011-12

(Rs. crore)

Sector	Year (2010-11)	Year (2011-12)
Primary	8,22,415	8,47,744
Secondary	12,84,941	13,34,249

Tertiary	**28,29,650**	**30,61,589**
Total	**49,37,006**	**52,43,582**

Source: Economic Survey, 2012-13

Solution:

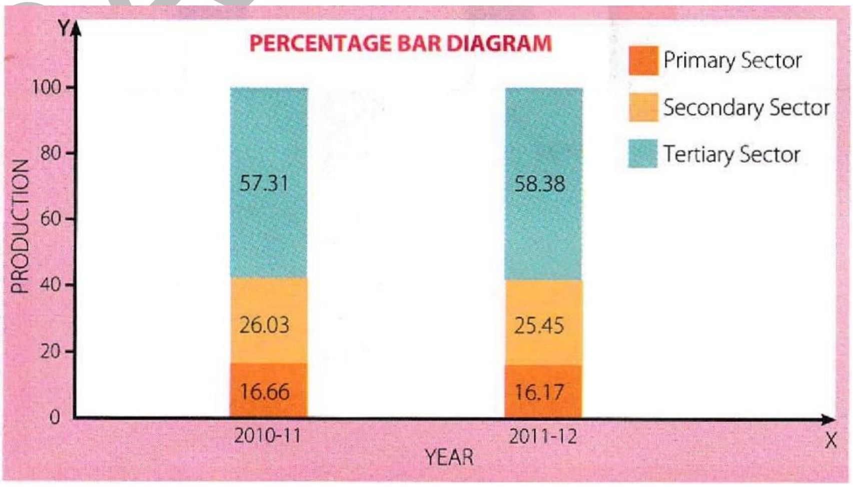
These data are first reduced to percentages and then presented in terms of percentage bar diagram, as under:

Percentage Contribution of Different Sectors in National Income

Sector	**Year (2010-11)**	**Cumulative Percentage**	**Year (2011-12)**	**Cumulative Percentage**
Primary	16.66	16.66	16.17	16.17
Secondary	26.03	42.69	25.45	41.62
Tertiary	57.31	100	58.38	100
Total	**100**		**100**	

Gross Domestic Product by Industry of Origin

in 2010-11 and 2011-12



(5) **Deviation Bar Diagrams:** The deviation bar diagrams are used to compare the net deviation of related variables with respect to time and location. Bars representing

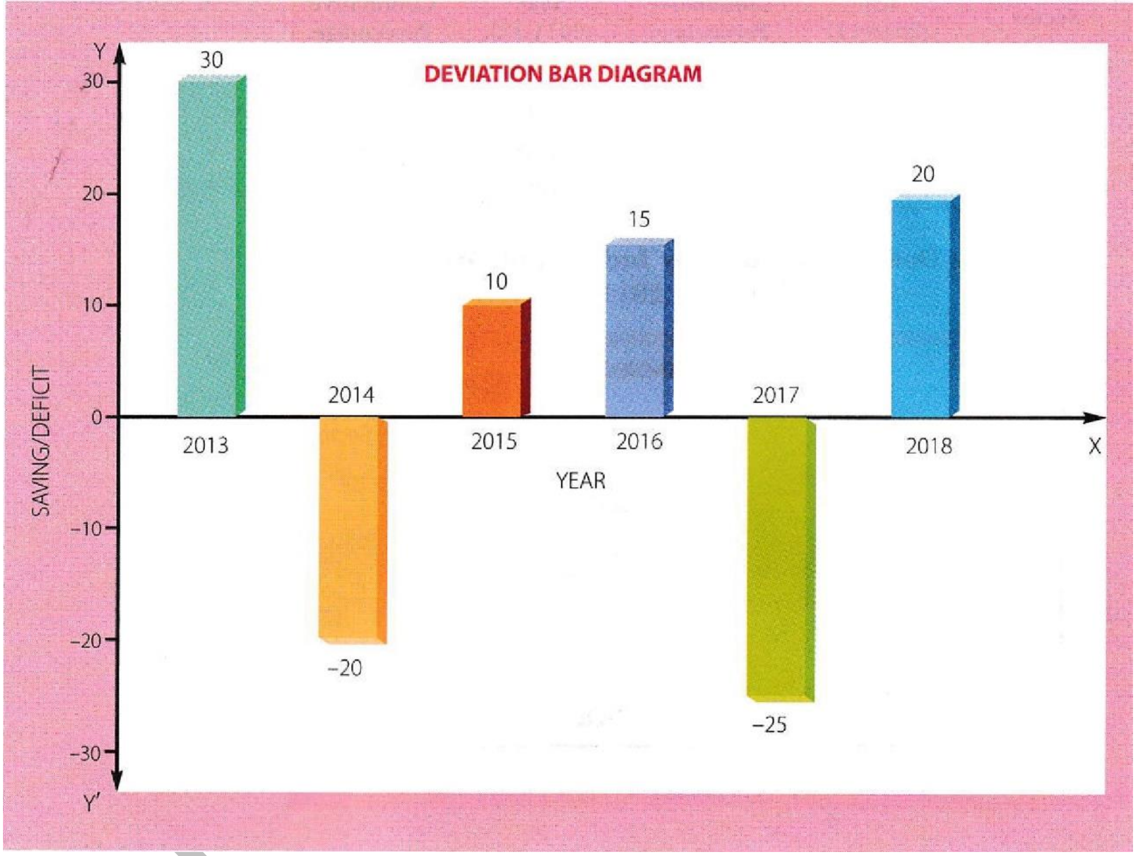
positive and negative deviations are drawn above and below the base line. Such type of diagrams represents the deviations in magnitude as well as in direction.

Illustration.

Represent the following data by a deviation bar diagram:

Year	2013	2014	2015	2016	2017	2018
Saving/Deficit (in '000 Rs.)	30	-20	10	15	-25	20

Solution:



Saving/Deficit 2013 to 2018 (in '000 Rs.)

2. PIE OR CIRCULAR DIAGRAMS

Pie diagram is a circle divided into various segments showing the per cent values of a series. This diagram does not show absolute values. Construction of a pie diagram involves the following steps:

(i) As a first step, absolute values of the series are converted into per cent values. To illustrate, if in a class of 45 students, 20 are first divisioners, 15 are second divisioners and 10 are third divisioners, then in tennis of per cent values it would mean $\frac{20}{45} \times 100$

or 44.45% are first divisioners, $\frac{15}{45} \times 100$ or 33.33% are second divisioners, and $\frac{10}{45} \times 100$ or 22.22% are third divisioners.

(ii) Second, draw a circle and note the fact that within a circle we can draw 4 angles of 90° each (See the adjacent figure), so that a circle comprises 90° × 4 = 360°. ‘

$$\angle abc + \angle abd + \angle ebc + \angle ebd = 360^{\circ}$$

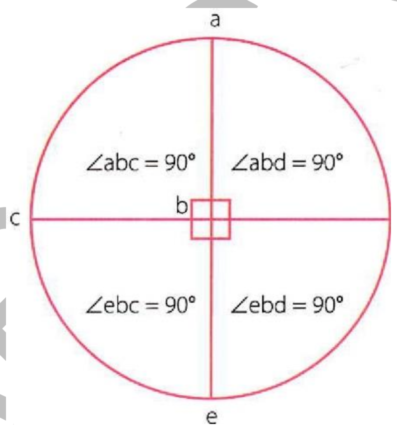
However, the circle called PIE, is to be finally divided on the basis of the given/actual data. Thus, given the percentage share of first divisioners, second divisioners and third divisioners as 44.45, 33.33 and 22.22 respectively (adding up to 100) we divide the circle in terms of different angles as under:

$$\frac{44.45}{100} \times 360^{\circ} = 160^{\circ}$$

$$\frac{33.33}{100} \times 360^{\circ} = 120^{\circ}$$

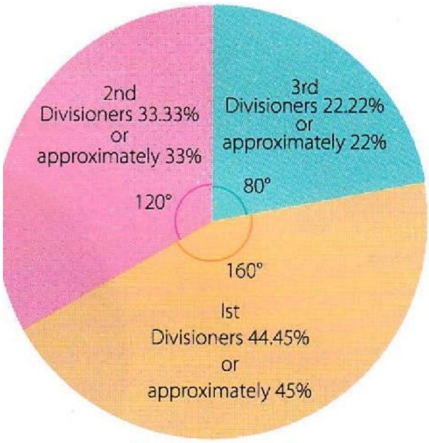
$$\frac{22.22}{100} \times 360^{\circ} = 80^{\circ}$$

∴ 160° + 120° + 80° = 360°



Accordingly, 44.45% of the First divisioners would mean 160° out of 360°, 33.33% of the Second divisioners would mean 120° out of 360°, and 22.22% of Third divisioners would mean 80° out of 360° of the circle.

(iii) Show each value in the circle clockwise as shown in Pie diagram-1. This indicates percentage distribution of the Students of Class XI according to their division status.



Pie Diagram–1

Illustration.

In 2011-12, Net Domestic Product by Industry of origin (at 2004-05) is as given below. Present this information in the form of a pie diagram.

Sector	% Share
Primary	16,2
Secondary	25.4
Transport	27.5
Finance and Insurance	18.1
Community and Social Services	12.8
Total	100

Source: Economic Survey, 2012-13

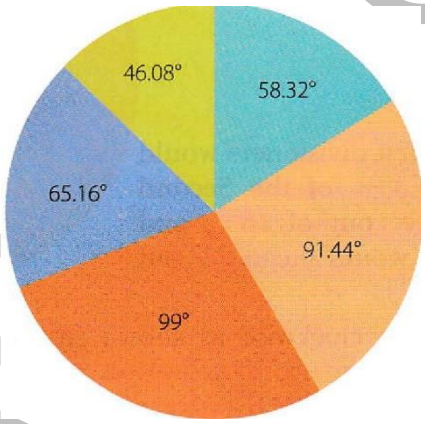
Solution:

Percentage value are converted into component parts of 360° of a circle, and then shown as different components of the circle.

Sector	% Share	Share in terms of Components of 360°
Primary	16.2	$\frac{16.2}{100} \times 360^{\circ} = 58.32^{\circ}$

Secondary	25.4	$\frac{25.4}{100} \times 360^{\circ} = 91.44^{\circ}$
Transport	27.5	$\frac{27.5}{100} \times 360^{\circ} = 99^{\circ}$
Finance and Insurance	18.1	$\frac{18.1}{100} \times 360^{\circ} = 65.16^{\circ}$
Community and Social Services	12.8	$\frac{12.8}{100} \times 360^{\circ} = 46.08^{\circ}$

Gross Domestic Product according to the Industry of Origin



Pie Diagram-2

Secondary

Transport

Finance and insurance

Community and Social Services

Multiple Choice Questions

Select the correct alternative:

1. Which of the following refer to geometric form of data presentation?

(a) Bar diagrams

(b) Histogram

(c) Pie diagrams

(d) Both (a) and (c)

2. The other name of pie diagram is:

(a) circular diagram

(b) bar diagram

(c) histogram

(d) polygon

3. Diagrams which show total values of a set of data simultaneously are known as:

(a) percentage bar diagrams

(b) differential bar diagrams

(c) deviation bar diagrams

(d) multiple bar diagrams

4. Which of the following statements is correct?

(a) Bars may be vertical or horizontal

(b) Breadth of the bars remains the same

(c) All bars are based on some common base line

(d) All of these

5. Diagrams which are used to compare the net deviation of related variables with respect to time and location are:

(a) deviation bar diagrams

(b) simple bar diagrams

(c) multiple bar diagrams

(d) pie diagrams

6. The most attractive method of data presentation is:

(a) Diagrammatic

(b) Textual

(c) Tabular

(d) Either (a) or (b)

7. In a bar diagram, the bars are:

(a) Horizontal

(b) Vertical

(c) Either (a) or (b)

(d) None of the above

8. Diagrammatic representation of data is done by:

(a) Pictures

(b) Charts

(c) Diagrams

(d) All these

9. Sub-divided bar diagram is used to:

(a) Study relation between different components

(b) Compare different components of a variable

(c) Either (a) or (b)

(d) Both (a) and (b)

10. The most appropriate diagram to represent the data relating to the monthly expenditure on different items by a family is:

(a) Histogram

(b) Pie diagram

(c) Frequency polygon

(d) Line graph

11. When for some countries, the magnitudes are small and for other, the magnitudes are very large, to portray the data, it is preferred to construct:

(a) Deviation bar diagram

(b) Duo-directional bar diagram

(c) Broken-Scale bar diagram

(d) Any of the above

12. The most accurate mode of data presentation is:

(a) Diagrammatic method

(b) Tabulation

(c) Textual presentation

(d) None of these

13. Details are shown by:

(a) Charts

(b) Tabular presentation

(C) Both (a) and (b)

(d) None of these

14. In tabulation, source of the data, if any, is shown in the:

(a) Source Note

(b) Body

(C) Stub

(d) Caption

15. The headings of the rows given in the first column of a table are called:

(a) Stubs

(b) Titles

(c) Captions

(d) Prefatory notes

16. The column heading of a table are known as:

(a) Stubs

(b) Sub-titles

(c) Reference notes

(d) Captions

17. For tabulation, ‘caption’ is:

(a) The lower part of the table

(b) The main part of the table

(c) The upper part of the table

(d) The upper part of a table that describes the column and sub-column

18. Which of the following statements is not true for tabulation`

(a) Complicated data can be presented

(b) Facilitates comparison between rows and not columns

(c) Statistical analysis of data requires tabulation

(d) Diagrammatic representation of data requires tabulation

19. 'Stub' of a table is the

(a) Right part of the table describing the columns

(b) Left part of the table describing the columns

(c) Right part of the table describing the rows

(d) Left part of the table describing the rows

20. Whether classification is done first or tabulation`

(a) Classification follows tabulation

(b) Classification precedes tabulation

(c) Both are done simultaneously

(d) No criterion

Answers

Multiple Choice Questions

1. (d)	**2. (a)**	**3. (b)**	**4. (d)**
5. (a)	**6. (a)**	**7. (c)**	**8. (d)**

9. (d)	**10. (b)**	**11. (c)**	**12. (b)**
13. (b)	**14. (a)**	**15. (a)**	**16. (d)**
17. (d)	**18. (b)**	**19. (d)**	**20. (b)**

CHAPTER 7

FREQUENCY DIAGRAMS— HISTOGRAM, POLYGON AND OGIVE

Frequency Diagrams relate to diagrammatic presentation of 'frequency distributions. In such series, values of a variable repeat themselves number of times. That is, different values of a variable happen to occur different number of times. Like, for example, corresponding to the range of marks between 40-50, there are 50 students, corresponding to the range of marks between 50-60 there are 40 students; corresponding to the range of marks between 60-70 there are 20 students, and so on, there are three important forms of frequency diagrams, viz.,

(i) Histogram,

(ii) Polygon, and

(iii) Ogive.

1. HISTOGRAM

A Histogram is a graphical presentation of a frequency distribution of a continuous series. While constructing a Histogram, values of the variable are shown on the X-axis, and their frequencies on the Y-axis. Frequencies (corresponding to different class intervals) are shown by rectangles. The width of the rectangles are according to the width of the corresponding class intervals. Height of the rectangles is in proportion to the frequencies of the class intervals. Different rectangles meet each other from left to right. If the data are in the form of an inclusive series, these are first converted into exclusive series. It may be noted that Histograms are different from Bar diagrams because these are two-dimensional diagrams. Both length and breadth of the rectangles are considered for purpose of a comparison. Bar diagrams are only one-dimensional where only length of the bars is considered for purpose of a comparison.

If frequencies are expressed in terms of percentage in the distribution, then the same are shown as percentage on the graph instead of the number of items.

Histograms are drawn only when data are in the form of grouped frequency distribution or when the data are in the form of frequency distribution of continuous series. Histograms are never drawn for a discrete variable or for a data set making a discrete series.

Histograms of Frequency Distribution are of two types:

(1) Histograms of Equal Class Intervals, and

(2) Histograms of Unequal Class Intervals.

(1) Histograms of Equal **Class Intervals:** Histograms of equal class intervals are those which are based on the data with equal class intervals. A series with equal class intervals would make a histogram including rectangles of equal width. Length of the rectangles would be different in proportion to the frequencies of the class intervals.

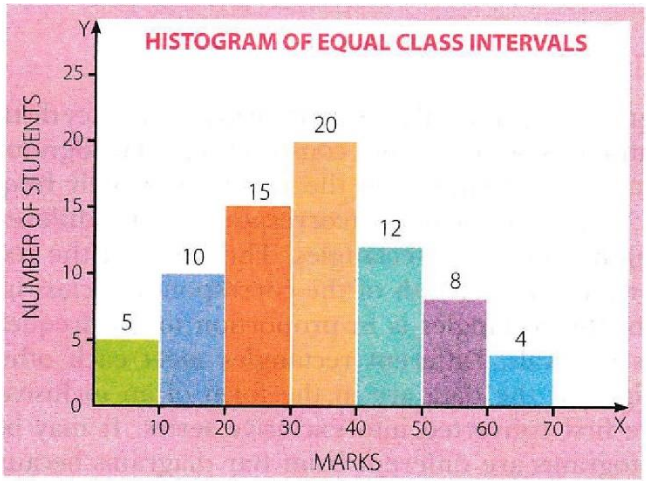
Illustration.

The table below shows number of students of a college corresponding to different range of marks in Statistics. Present the information in the form of a histogram.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of Students	5	10	15	20	12	8	4

Solution:

Number of Students Corresponding to Different Range of Marks in Statistics



(2) Histograms of Unequal Class Intervals: A histogram of unequal class interval is the one which is based on the data with unequal class intervals. When the data of class intervals are unequal, width of the rectangles would be different. The width of the rectangles would increase or decrease depending upon the increase or decrease in the size of the class intervals. Before presenting the data in the form of graphs, frequencies of unequal class-intervals are adjusted. First, we note a class of the smallest intervals. Other classes are noted in the increasing order of their class-intervals. If the size of one class interval is twice the smallest size in the series, frequency of that class is divided by ‘2’. Likewise, if the size of a class interval is three times the size of the smallest class interval in the series, frequency of that class is divided by ‘3’ and so on.

Adjustment Factor for any Class = $\frac{\textbf{Class Interval of the Concerned Class}}{\textbf{Lowest Class Interval}}$

To illustrate, suppose the class with the smallest intervals is 5-10, and the class with the largest interval is 10-20 the frequency of which is 12. Here the class interval of the bigger class is 10 which is twice as much as the size of the class interval of the smallest class, i.e., 5. The bigger class interval is divided in two parts 10-15 and 15-20 and accordingly, the frequency of the bigger class, 12 would be divided by 2, that is 12 ÷ 2 = 6. The divider 2 in this case is called Adjustment Factor.

Illustration.

Present the following data in the form of a histogram:

Weekly Wages (Rs.)	Number of Workers
10-15	7
15-20	10
20-25	27
25-30	15
30-40	12
40-60	12
60-80	8

Solution:

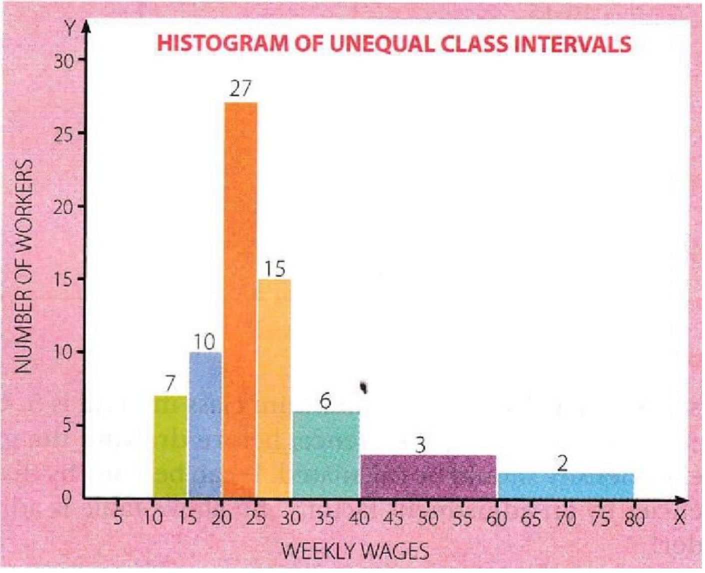
In this frequency distribution minimum class interval is 5! Other class intervals are 10 and 20. Hence, before drawing the graph, frequency density should be calculated. It can be done by dividing frequencies by an adjustment factor. The above table is adjusted as under:

Adjustment of Frequencies of Unequal Class Intervals

Weekly Wages (Rs.)	Number of Workers	Adjustment Factor	Frequency Density
10-15	7	$\frac{5}{5}=1$	7÷1=7
15-20	10	$\frac{5}{5}=1$	10÷1=10
20-25	27	$\frac{5}{5}=1$	27 ÷ 1 = 27
25-30	15	$\frac{5}{5}=1$	15 ÷1 = 15
30-40	12	$\frac{10}{5}=2$	12÷2 = 6
40-60	12	$\frac{20}{5}=4$	12 ÷ 4 = 3
60-80	8	$\frac{20}{5}=4$	8÷4 = 2

In the above table, the class interval for the first four classes is 5. Fifth class, however, is of the interval of 10 (40 - 30 = 10) which is twice as much as the class interval of the first four classes. Accordingly, frequency of the fifth class is divided by 2. Further, class interval of the 6th class is 20 (60 - 40 = 20) which is 4 times the minimum class interval of 5. Accordingly, frequency of this class is divided by 4. Likewise, the frequencies of other classes have been adjusted. It is based on this table that the histogram is drawn, as given below:

Number of Workers Corresponding to Different Range of Weekly Wages



2. POLYGON

Polygon is another form of diagrammatic presentation of data. It is formed by joining mid-points of the tops of all rectangles in a histogram. However, a polygon can be drawn even without constructing a histogram. For this, mid-values of the classes of a frequency distribution are marked on X-axis of the graph; the corresponding frequencies are marked on the Y-axis. Using a foot rule, all points indicating frequencies of the different classes are joined to make a graph, called frequency polygon. Both the sides of the frequency polygon are extended to meet the X-axis, at the mid-points of the immediately lower or higher imagined class intervals of zero frequency. This is done to ensure that the area of a frequency polygon is the same as that of the corresponding histogram.

Illustration.

Following table shows number of students of a college corresponding to different range of marks in Statistics. Make a frequency polygon.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of Students	5	10	15	20	12	8	5

Solution:

A frequency polygon of the above data is presented in the following graph. In this graph, data have been first presented in the form of a histogram. Mid-points of the tops of rectangles in the Histogram have been marked and then joined using a scale or foot rule. The end-points of the polygon be joined to the immediate lower or higher mid-points (as the case may be) at zero frequency with the base-line. It is done to equate the area of polygon with the area of histogram. Here is an illustration of constructing a frequency polygon by making a histogram.

What is the difference between a Histogram and a Frequency Polygon?

A histogram becomes a frequency polygon if we draw a line joining mid-points of the tops of all rectangles in a Histogram. It is important that the midpoints are joined using a foot-rule to make a straight line.

Illustration.

Present the following data in the form of frequency polygon:

Marks	10-20	20-30	30-40	40-50	50-60	60-70
Number of Students	10	13	20	22	15	10

Solution:

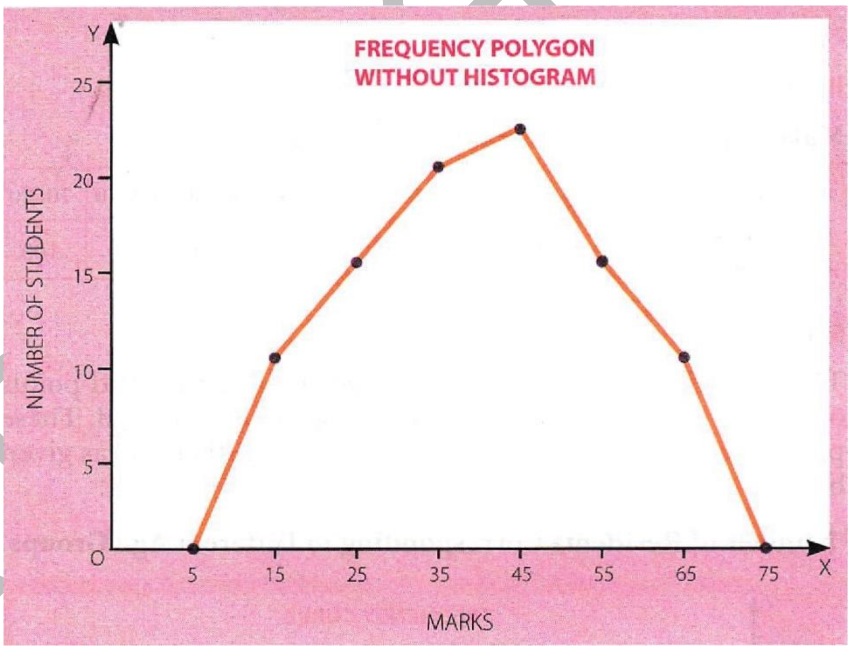
To construct a frequency polygon without first constructing a histogram, one should first mark mid-points of the class intervals on the X-axis. Frequencies are then marked corresponding to those mid-points.

Different points of frequencies are joined to make a frequency polygon.

Marks	Mid-value	Number of Students
10-20	$\frac{10+20}{2} = 15$	10
20-30	$\frac{20+30}{2} = 25$	15
30-40	$\frac{30+40^{\circ}}{2} = 85$	20
40-50	$\frac{40+50}{2} = 45$	22
50-60	$\frac{50+60}{2} = 55$	15
60-70	$\frac{60+70}{2} = 65$	10

Based on these data, frequency polygon has been drawn without constructing a histogram, as given below.

Number of Students Corresponding to Different Range of Marks



3. FREQUENCY CURVE

It is just a variant of polygon. A Frequency Curve is a curve which is plotted by joining the mid-points of all tops of a histogram by freehand smoothed curves and not by straight lines. Area of a frequency curve is equal to the area of a histogram or frequency polygon of a given data set. While drawing a frequency curve, we should eliminate angularity of the polygon. Accordingly, points of a frequency polygon are joined through a freehand smoothed curve rather than straight lines.

The basic difference between a Frequency Polygon and a Frequency Curve

Both frequency polygon and frequency curve are drawn by joining the mid-points of ail tops of **a** histogram. But in case of frequency polygon the points are joined using a foot-rule (to make a i straight line), whereas in case of frequency curve the points are joined using **a** freehand.

Illustration.

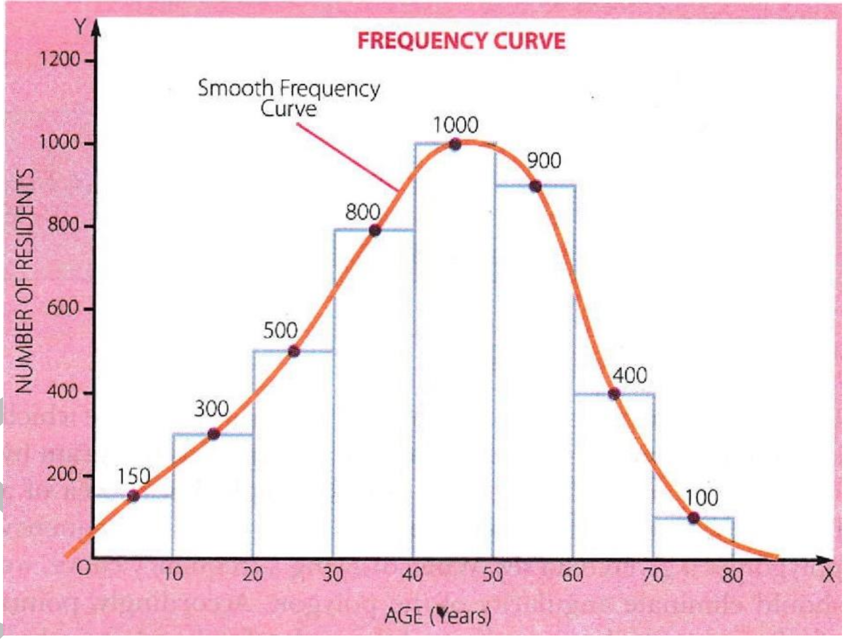
Make a frequency curve of the following data:

Age (Years)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of Residents	150	300	500	800	1,000	900	400	100

Solution:

The given data set is first converted into a histogram. Mid-points of the top of the rectangles of the histogram are marked. These points are joined through a freehand smoothed curve, as given below:

Number of Residents Corresponding to Different Age Groups



4. OGIVE OR CUMULATIVE FREQUENCY CURVE

Ogive or Cumulative Frequency Curve is the curve which is constructed by plotting cumulative frequency data on the graph paper; in the form of a smooth curve.

A cumulative frequency curve or ogive may be constructed in two ways:

(1) Less than Method: In this method, beginning from upper limit of the 1st class interval we go on adding the frequencies corresponding to every next upper limit of the series. Thus in a series showing 0-5, 5-10 and 10-15 as different class intervals, we will find the frequency for less than 5, for less than 10 and for less than 15. The frequencies are added up to make Less than Ogive’.

(2) **More than Method:** In this method, we take cumulative total of the frequencies beginning with lower limit of the 1st class interval. Thus, in a series showing 0-5, 5-10 and 10-15 as different class intervals, we find the frequency for more than 0, for more than 5 and for more than 10. The frequencies thus presented make a 'More than Ogive’.

The basic difference between 'Less than' and 'More than' Ogives

In less than ogives, frequencies are added starting from the upper limit of the 1st class interval of the frequency distribution. On the other hand, in case of 'more than' ogives, frequencies are added starting from the lower limit of the 1st class interval of the frequency distribution. Accordingly, while in case of less than' ogive the cumulative total tends to increase, in case of ‘more than' ogive, the cumulative total tends to decrease.

Illustration.

Following data relate to the marks secured by students in their Statistics paper. Graph these data in the form of less than ogive and more than ogive.

Marks	**0-5**	5-10	**10—15**	**15-20**	**20-25**	**25-30**	**30-35**	**35-40**
Number of Students	4	6	**10**	**10**	**25**	**22**	**18**	**5**

Solution:

First we construct a cumulative frequency table of ‘less than’ and ‘more than’ type, and then draw the graphs, as under:

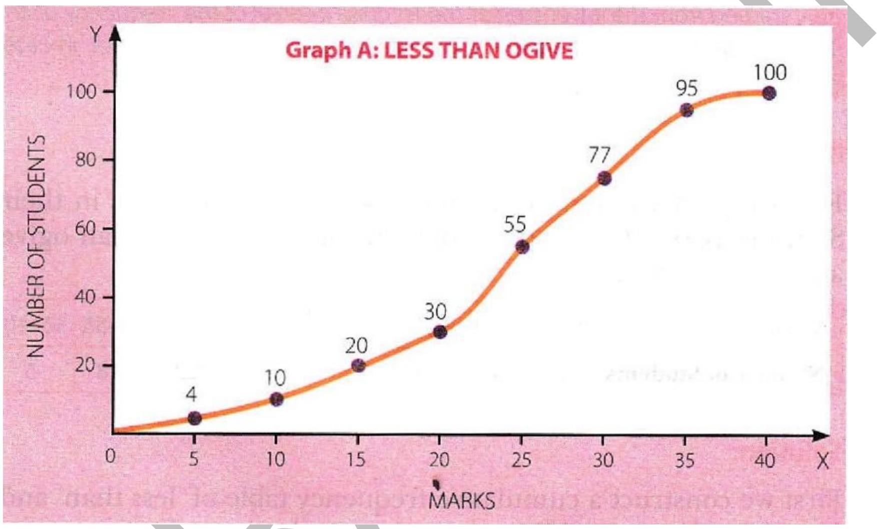
Cumulative Frequency Table

Less than Type		**More than Type**	
Marks **(1)**	**Cumulative** **Frequencies** **(2)**	**Marks** **(3)**	**Cumulative** **Frequencies** **(4)**
Less than 5	4	More than 0	100
Less than 10	4 + 6 = 10	More than 5	100- 4 = 96
Less than 15	10+10= 20	More than 10	96 - 6 = 90
Less than 20	20+10= 30	More than 15	90- 10 = 80
Less than 25	30 +25 = 55	More than 20	80-10 = 70

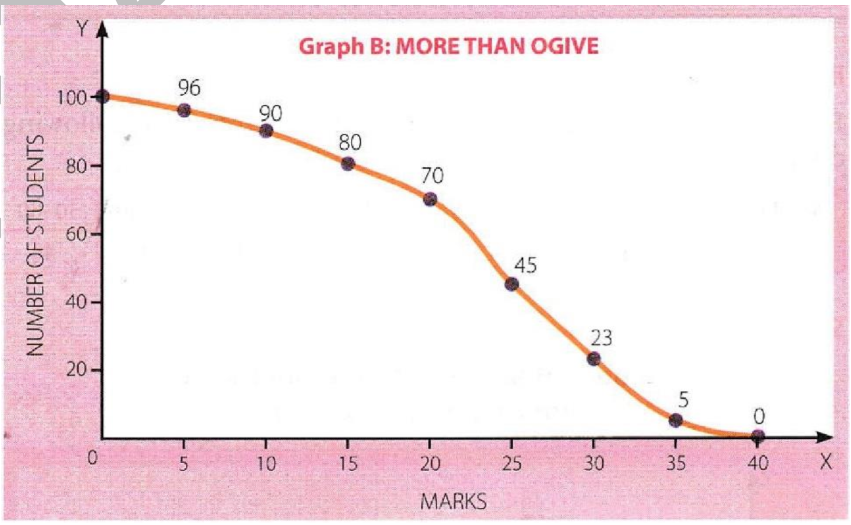
Less than 30	55 +22 = 77	More than 25	70 - 25 = 45
Less than 35	77+18= 95	More than 30	45 - 22 = 23
Less than 40	95 + 5 = 100	More than 35	23-18= 5
		More than 40	5-5=0

Graph A shows 'less than' cumulative frequencies, Graph B shows 'more than' cumulative frequencies, and Graph C shows both the 'less than' as well as 'more than' frequencies simultaneously.

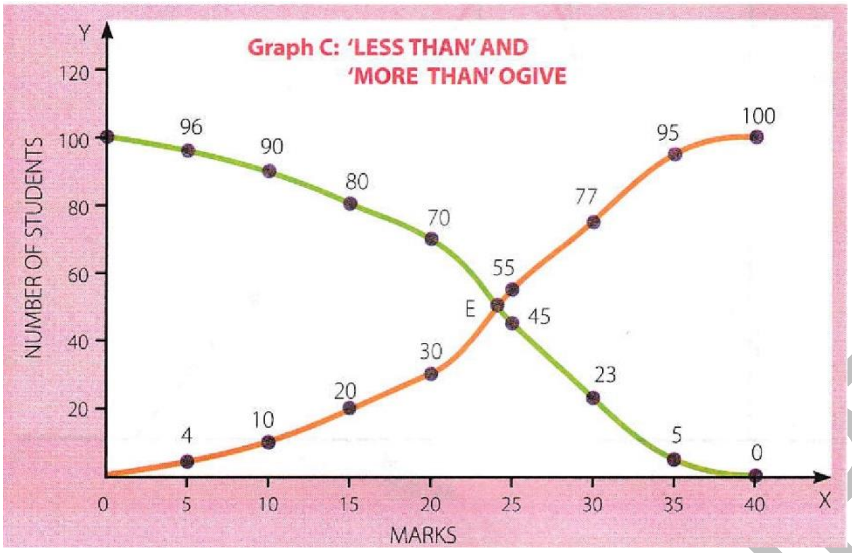
Students obtaining Marks in Statistics



Students obtaining Marks in Statistics



Students obtaining Marks in Statistics

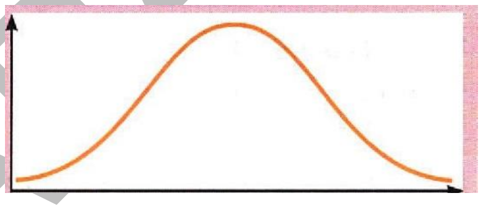


DIFFERENT SHAPES OF FREQUENCY CURVES

Frequency curves are of different shapes. These shapes indicate the nature of frequency distributions of different data-sets. Some of the important shapes of frequency curves are as under:

(1) Normal Curve or Symmetrical Curve

Graph-A shows a normal curve or symmetrical curve. In such curves, frequencies tend to increase gradually, followed by the tendency to stabilise and finally the tendency to decline. The highest frequency in such curves lies at the top center of the curve. The highest frequency, decreases on its either side at the uniform rate.



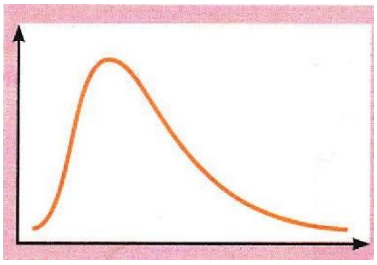
Graph-A: Symmetrical of Bell Shaped

This type of curve is symmetrical about a line which divides the curve in two equal parts. Such curves are also called Bell shaped. In practical life normal curves are very rare. However, these are of immense significance in the context of statistical analysis. Normal curves are extremely useful in understanding problems relating to sampling of data.

(2) Positive Skewed Curve

This is one of the asymmetrical curves or skewed curves.

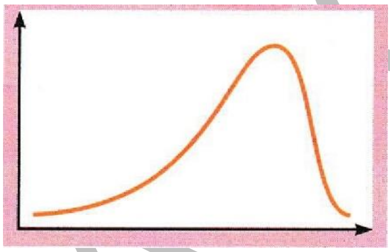
Asymmetry or Skewness is that feature of a statistical distribution which indicates that, on both sides of the highest frequency mark, the frequencies do not decrease at the uniform rate. The rate of decrease on one side is less or more as compared to the other side. Accordingly, the one side of the curve is more spread as compared to the other. Curves which are skewed to the right (Graph-B) have their right side more spread than the left one. That is, these are tailed to the right. Obviously, this means that lower values (or initial classes) in such distributions have greater frequencies and higher values in the distribution have smaller frequencies.



Graph-B: Skewed to the right (positive skewness)

(3) Negative Skewed Curve

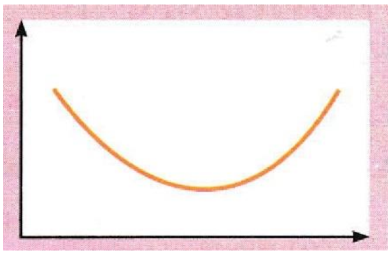
This is another kind of an asymmetrical curve or skewed curve. The only difference between this and the positive skewed curve is that this curve is skewed to the left. That is, such curves are tailed to the left (Graph-C). It means that lower values (or initial classes) in such distributions have smaller frequencies and higher values in the distribution have greater frequencies.



Graph-C: Skewed to the left (negative skewness)

(4) U-Shaped Curve

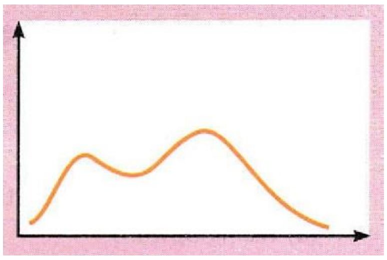
U-shaped curves are formed if there are two high points in a series both having equal or nearly equal frequency, at the lowest and highest values (Graph-D). Most of the frequencies of a distribution are thus concentrated around the lowest and highest-class intervals. There are only few frequencies marks corresponding to middle class intervals.



Graph-D: U-Shaped

(5) Bi-modal Curve

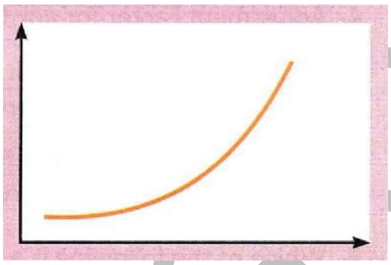
Graph-E shows a Si-modal curve. This curve is drawn when, in a series, there are two classes with highest frequencies. In such series, there is no unique modal value. That is why the graphic representation of such series make a Bi-modal curve.



Graph-E: Bi-modal

(6) J-Shaped Curve

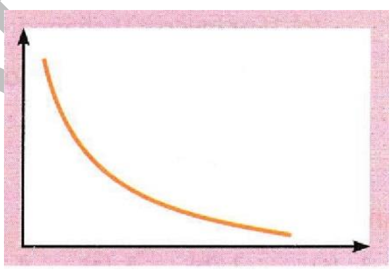
Graph-F shows a J-shaped curve. This is drawn when, in a distribution, frequencies tend to increase with each class interval. In other words the curve moves from the low frequencies to the high frequencies. Thus, highest frequencies are found around the upper end of the curve. Such curves are shaped like the letter T in English language.



Graph-F: J-Shaped

(7) Reverse J-Shaped Curve

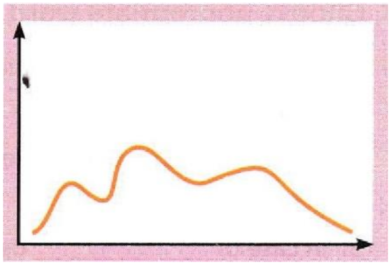
Graph-G shows a reverse J-shaped curve. This type of curve is drawn when there are highest frequencies corresponding to lowest values in the distribution. Conversely, there are lowest frequencies corresponding to highest values in the distribution.



Graph-G: Reverse J-Shaped

(8) Mixed Curve or Multi-modal Curve

These curves are formed when there is no specific pattern of the frequencies corresponding to different values (or classes) in the given distribution of data. That is, frequencies keep on increasing and decreasing corresponding to different classes in the data set. Graph-H shows a mixed curve, also called a multi-modal curve.



Graph-H: Mixed Curve

Multiple Choice Questions

Select the correct alternative:

1. A Histogram is a graphical presentation of a frequency distribution of a:

(a) individual series

(b) discrete series

(c) continuous series

(d) none of these

2. Which of the following diagrams is drawn by joining mid-points of the tops of all rectangles in a histogram:

(a) frequency distribution

(b) frequency polygon

(c) frequency curve

(d) none of these

3. What is the shape of 'less than ogive'?

(a) Rising upward

(b) Falling downward

(c) Parallel to X-axis

(d) Parallel to Y-axis

4. Adjustment Factor for any Class is equal to:

(a) $\frac{\text{Class Interval of the First Class}}{\text{Lowest Class Interval}}$

(b) $\frac{\text{Class Interval of the Concerned Class}}{\text{Lowest Class interval}}$

(c) $\frac{\text{Class Interval of the First Class}}{\text{Upper Class Interval}}$

(d) $\frac{\text{Class Interval of the Concerned Class}}{\text{Upper Class Interval}}$

5. Which of the following is a shape of frequency distribution curve?

(a) A-shaped

(b) B-shaped

(c) U or inverse U-shaped

(d) All of these

6. Normal curves are also known as:

(a) J-shaped curves

(b) L-Shaped curves

(c) U-shaped curves

(d) bell-shaped curves

7. Histogram is prepared in case of:

(a) Individual Series

(b) Discrete Series

(c) Continuous Series

(d) neither of the above

8. Graphs can be useful for:

(a) Showing trends in data

(b) Adding visual appeal to business reports

(c) Making comparisons

(d) All of the above

9. A comparison among the class frequencies is possible only in

(a) Ogive

(b) Histogram

(c) Frequency polygon

(d) Either (a) or (b)

10. The breadth of the rectangle is equal to the length of the class-interval in

(a) Ogive

(b) Histogram

(c) Both (a) and (b)

(d) None of these

11. A frequency polygon is obtained by:

(a) Constructing a frequency histogram

(b) Constructing a cumulative frequency histogram

(c) Linking mid-points from a frequency histogram

(d) Using a line graph

12. While preparing time series graph, we show ___________ on the X-axis:

(a) Time

(b) income

(c) Expense

(d) all the above

13. Ogive curve occur for,

(a) More than type distribution

(b) Less than type distribution

(c) Both (a) and (b)

(d) None of (a) and (b)

14. It is always possible to construct a histogram from the:

(a) Data

(b) Frequency polygon

(c) Both from the data and frequency polygon

(d) None of these

15. Mode is found graphically by:

(a) Frequency Polygon

(b) Ogive

(c) Histogram

(d) None of these

16. To find the number of observations less than any given value, we use:

(a) Grouped frequency distribution

(b) Single frequency distribution

(c) Cumulative frequency distribution

(d) None of these

17. A simple frequency distribution of a discrete variable is represented by:

(a) Ogive

(b) Histogram

(c) Frequency Polygon

(d) None of these

18. Median of a distribution can be obtained from

(a) Histogram

(b) Frequency polygon

(c) Less than type ogives

(d) None of these

19. Diagrammatic representation of the cumulative frequency distribution is:

(a) Frequency Polygon

(b) Ogive

(c) Histogram

(d) None of these

20. Graph of successive points of a distribution joined by straight lines in statistical terminology is known as:

(a) Frequency distribution

(b) Frequency polygon

(c) Trend

(d) Cumulative distribution curve

21. If you are interested in how earnings of a company have fluctuated over time, it would be best to use:

(a) Bar graph

(b) Time series graph

(c) Pie chart

(d) Histogram

22. When the values are large in magnitude in a chronological series and variation amongst values is small, a graph is better drawn by choosing:

(a) A false base line

(b) Wide scale

(C) Narrow scale

(d) None of these

Answers

Multiple Choice Questions

1. (c)	**2. (b)**	**3. (a)**	**4. (b)**	**5. (c)**	**6. (d)**
7. (c)	**8. (d)**	**9. (b)**	**10. (b)**	**11. (c)**	**12. (a)**
13. (c)	**14. (c)**	**15. (c)**	**16. (c)**	**17. (c)**	**18. (c)**
19. (b)	**20. (b)**	**21. (b)**	**22. (a)**		

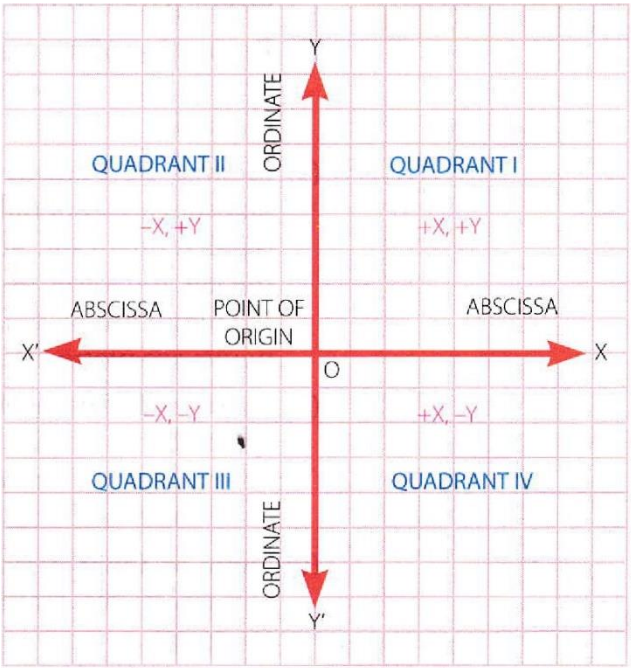
CHAPTER 8

ARITHMETIC LINE-GRAPHS OR TIME SERIES GRAPHS

When a set of statistical data are presented on a graph paper, it is called a graph. Presenting the data on a graph paper, we get different points, each point corresponding to a value of a statistical series. By joining the points, we get a line which shows how a variable tends to change. Each point on the line corresponds to an arithmetic value of the variable under study like, for example, the value of exports or imports. Therefore, a graph showing arithmetic values of a variable (on a graph paper) is called ‘arithmetic line-graph’. Often such graphs are constructed to present time series data, that is, the data (like the value of exports or imports) corresponding to different weeks, months or the years. Therefore, arithmetic line-graphs are often called ‘time series graphs.

1. CONSTRUCTION OF A GRAPH

Keeping in mind the size and nature of data, a suitable intersecting point on the graph paper is assumed as a Point of Origin as point ‘O’ in the adjacent picture. Horizontal and vertical lines intersecting through this point are shown as bold lines. These bold lines are called Axis. Horizontal line from left to right is called Horizontal Axis, Abscissa or X-axis. The line going upward or downward is called Vertical Axis, Ordinate or Y-axis. The X-axis and Y-axis are mutually perpendicular



to each other. These axes divide the plain of the paper in four parts. Each part is called 'quadrant' as in the graph given on previous page.

Point of origin in the graph is indicated by the English letter 'O'. Positive values of X are taken towards right side on the horizontal axis and of Y towards upper side on the vertical line. Negative values of X are taken towards the left side on horizontal line and of Y towards the lower side on vertical line. Thus, in the first quadrant, both X and Y are positive, in the second, X is -ve while Y is +ve; in the third quadrant both X and Y are -ve; and in the fourth X is +ve while Y is -ve. In most cases, the data are positive figures. These are therefore presented only in the first quadrant.

Rules for Constructing a Graph

The following points must be kept in mind while constructing a graph:

(1) Heading: Every graph must have a suitable and precise heading. Heading must be self-explanatory about the nature of information in the graph.

(2) Choice of Scale: One should fix an appropriate scale on which data should be presented. An appropriate scale is the one by which the entire data are easily represented by the graph. The graph should be on the middle of the graph paper to make it attractive.

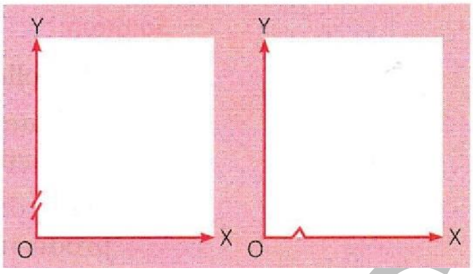
(3) **Proportion of Axis:** As far as possible, length of X-axis on the graph paper should be one and a half (I'/a) times the length of Y-axis.

(4) Method of Plotting the Points: Economics and business Statistics are generally positive. These are to be presented in the first quadrant. Accordingly, the point of origin is fixed to the left and lower portion of the graph paper. On the X-axis, the points are plotted from left to right and on the Y-axis, the points are plotted upward from bottom to top.

(5) Lines of Different Types: If more than one line or curve are to be drawn in the same graph, these lines should be differentiated from each other in the form of broken lines (------), dotted lines (......), bold lines (____), etc.

(6) Table **of Data:** It would be useful to give the table of data along with the graph of the data. This helps verification of the graph.

(7) Use of False Line: If the values in a series are very large and the difference between the smallest value and zero is high and if these values are to be indicated on Y-axis of the graph, then the Y-axis is started somewhere above the point 'O'.



(8) To Draw a Line or Curve: We mark different points on the graph paper corresponding to different values of a series. These points are joined to make a line or a curve. The joining line must be uniform throughout its length. It should not be of different thickness at its different points.

One Variable Graphs

One variable graphs are those graphs in which values of only one variable are shown with respect to some time period. Graphic presentation of the production of a factory between the months of January and June of a year, would be one variable graph.

Illustration.

Following table shows production of a factory between January and June. Present the information in the form of a one Variable time series graph.

Month	January	February	March	April	May	June
Production (Quintals)	5	7.5	5	10	12.5	15

Solution:

Take the following steps:

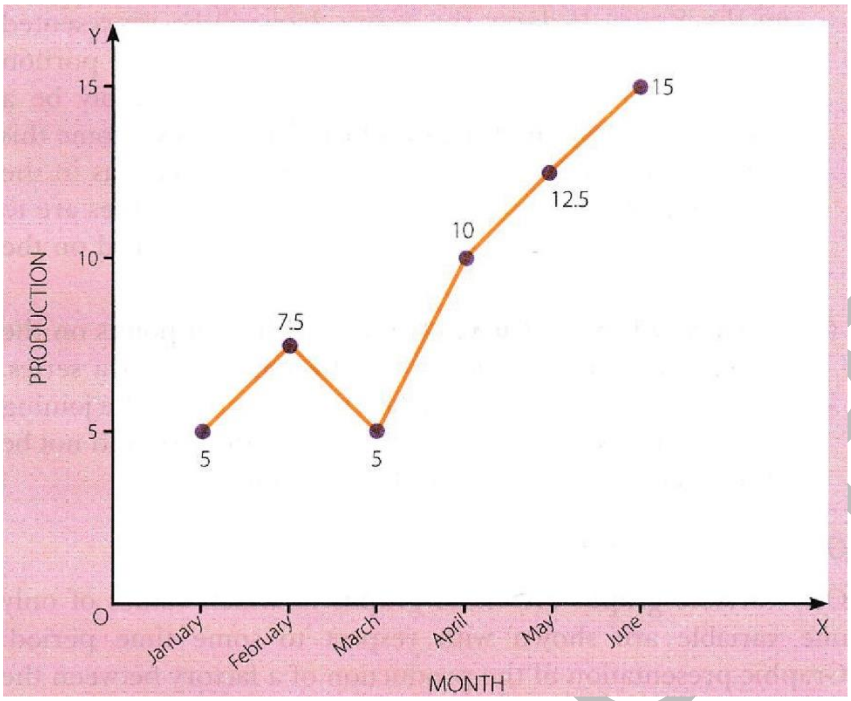
(i) Indicate time period (in terms of months) on the X-axis, and production on the Y-axis.

(ii) Mark different points on the graph indicating values of production corresponding to different months.

(iii) Join the points to get a graph showing the behaviour of production over time.

Monthly Production of a Factory

(Quintals)



Use of False Line: The use of false line is very common with respect to time series data. Following illustration explains how exactly it is done.

Illustration.

Following table gives hypothetical figures of exports from India during the years 2013-14 to 2017-18. Present the information in the form of a suitable graph.

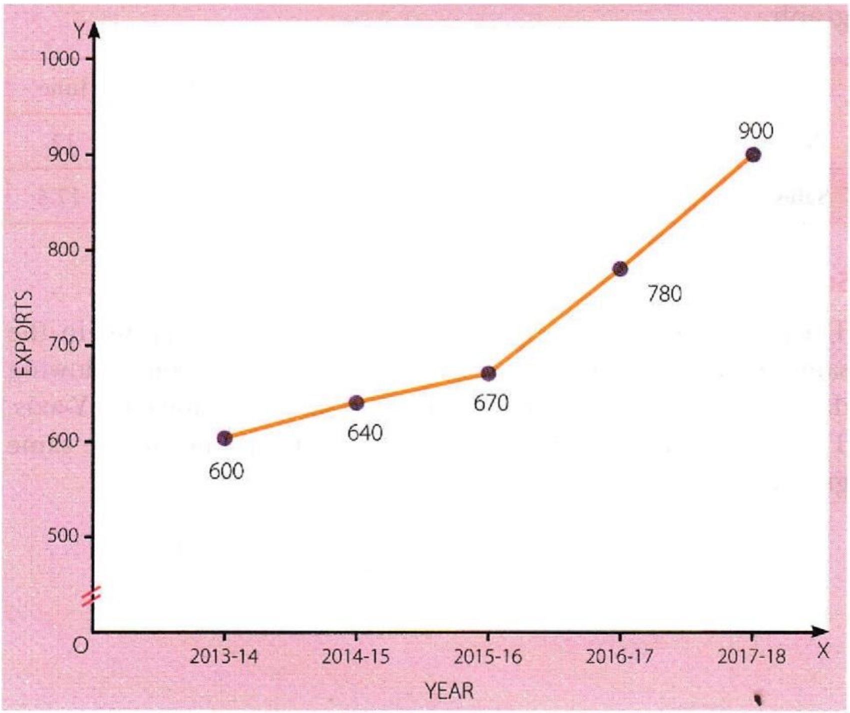
Year	**2013-14**	**2014-15**	**2015-16**	**2016-17**	**2017-18**
Exports (Rs. crore)	600	640	670	780	900

Solution:

Presentation of these data involves the construction of a false line. Because: (i) difference between the minimum export value and zero is very large, and (ii) various values of exports show large difference. Data are presented in a graph with a false base line as under:

Exports from India (2013-14 to 2017-18)

(Rs. crore)



Two or More than Two Variable Graphs

These are the graphs in which values of two (or more than two) variables are simultaneously shown with respect to some period of time. Data on the production and sale of a factory in different months would make a two-variable graph. Following illustration should make this point clear.

Illustration.

The following table gives data on the production and sales of a factory (in thousand rupees) between January and June. Present the information in the form of a two variable arithmetic-line graph.

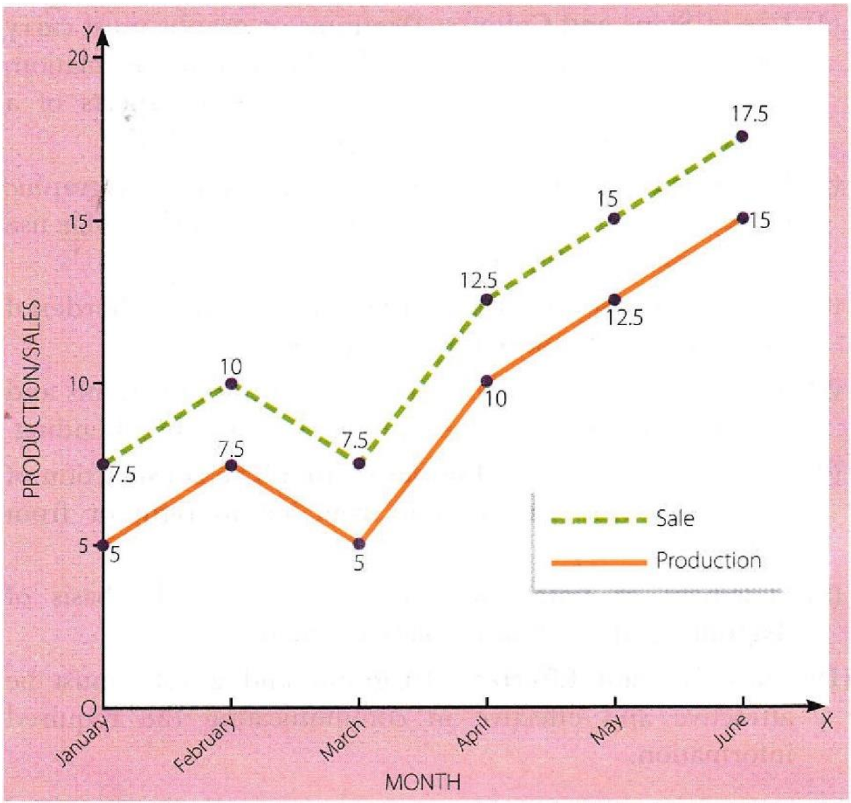
Month	January	February	March	April	May	June
Production	5	7.5	5	10	12.5	15
Sales	7.5	10	7.5	12.5	15	17.5

Solution:

These data will also be presented in the form of graph in the same manner as shown in the above graph. In graph, following data pertaining to both production and sales are shown on Y-axis. These are represented by two different graph lines in the same graph.

Monthly Production and Sales of a Factory

(000’ Rs.)



2. GENERAL RULES FOR CONSTRUCTING DIAGRAMS AND GRAPHS

Some of the general rules for constructing diagrams and graphs are as follows:

(1) Proper Size: Diagrams or graphs must suit the size of the paper. It should be neither too big nor too small.

(2) Proper Heading: Diagrams or graphs must bear proper heading. A heading must be simple, short and informative.

(3) Proper Scale: Before making a diagram/graph its scale should be properly determined and indicated.

(4) Use of Signs and Colours: Diagrams or graphs must carry some signs on the nature and classification of information. Colours may be used to indicate different aspects of a diagram. These signs and colours must be clarified.

(5) **Less Use of Words or Figures:** In diagrammatic or graphic presentation of data one should make minimum possible use of the words and figures.

(6) Drawing the Border: Diagrams or graphs must be bordered with bold lines to make them attractive.

(7) Simple: Simplicity is the principal feature of diagrams and graphs. These should not look to be complex and offending.

(8) **From Left to Right or Bottom to Top:** The construction of diagrams/graphs should flow from left to right or from bottom to the top.

(9) Statement of Data: Data which constitute the basis of diagrams/graphs should be clearly stated.

(10) Attractive and Effective: Diagrams and graphs must be attractive and effective in communicating the required information.

3. MERITS OF DIAGRAMMATIC AND GRAPHIC PRESENTATION

Notable merits of diagrammatic and graphic presentation are as under:

(1) Simple and Understandable Information: Even the most complex statistical information is made simple and understandable with the help of diagrams and graphs. One can understand the features of data merely by having a look at the picture.

(2) Lasting Impact: Diagrammatic or graphic presentation leaves a lasting impact on the reader’s mind. Information is not easily forgotten.

(3) No Need of Training or Specialised Knowledge: One needs no training or specialised knowledge in reading the diagrams and graphs. These are easily understood even by a layman.

(4) Attractive and Effective Means of Presentation: Diagrams and graphs are very attractive and effective means of presenting data. It is rightly said that a picture is worth of a thousand words.

(5) A Quick Comparative Glance: Diagrams/graphs facilitate a comparative glance at the data. Thus, data on investment in Private and Public sectors, when presented in the form of (say) bar diagrams, can be easily compared. One can easily note the broad differences between the two.

(6) Informative and Entertaining: Besides being informative, diagrammatic or graphic presentation is an entertaining means of data presentation. The beginners are just fascinated to draw pictures in the form of diagrams and graphs.

(7) Location of Averages: Using graphic technique, we can easily locate the values of certain averages, such as mode and median.

(8) Study of Correlation: Graphic presentation of data corresponding to different variables helps identify correlation between the variables. For example, if time series data on income and expenditure are plotted on the same graph, one is very likely to observe a very high degree of positive correlation between the two variables. Higher level of income is very likely to be associated with higher level of consumption, and vice versa.

4. LIMITATIONS OF DIAGRAMMATIC AND GRAPHIC PRESENTATION

Some of the limitations of diagrammatic and graphic presentation of statistical data are as follows:

(1) Limited Use: Only a limited set of data can be presented in the form of a diagram. In fact, diagrams and graphs are generally used only when comparisons are involved or when time-series data are to be presented.

(2) **Misuse:** Diagrams may be misused for false projection of the statistical facts, especially in case of advertisements. <

(3) **Only Preliminary Conclusions:** It may not be always easy to arrive at final conclusions after seeing the diagrams. Multiple information in the form of diagrams and graphs may offer only preliminary conclusions.

Multiple Choice Questions

Select the correct alternative:

1. Arithmetic line-graphs are also known as:

(a) linear graphs

(b) non-linear graphs

(c) time series graphs

(d) none of these

2. Axis divides the plain of a paper into:

(a) two quadrants

(b) three quadrants

(c) four quadrants

(d) none of these

3. In the first quadrant, the values of X and Y are:

(a) +ve

(b)-ve

(c) X is +ve and Y is -ve

(d) none of these

4. If the values in a series are very large and the difference between the smallest value and

zero is high, then we use _____________base line.

(a) original

(b) false

(c) true

(d) none of these

5. In which quadrant, the value of X will be positive but that of Y will be negative?

(a) 1st

(b) 2nd

(c) 3rd

(d) 4th

6. Graphs are always drawn with reference to:

(a) scale

(b) origin

(c) both (a) and (b)

(d) none of these

Answers

Multiple Choice Questions

1. (c)	**2. (c)**	**3. (a)**	**4. (b)**
5. (d)	**6. (a)**		

CHAPTER 9

MEASURES OF CENTRAL TENDENCY ARITHMETIC MEAN

1. CONCEPT AND DEFINITION OF CENTRAL TENDENCY

A Central Tendency refers to an average or a central value of a statistical series.

It is difficult for anyone to understand or remember a large group of raw data. One would like to know the critical value which represents all the items in a series. Such a value is called 'central tendency' or 'average value'. For instance it is very difficult to remember and understand the data concerning the income of millions of Indians. However, if it is said that in 2016-17, provisional estimates for average income of the people in India was Rs. 82,269 per annum, it will be easy for us to guess the economic condition of most of the Indians. It is this average value which is called central tendency of the series. It is also called measure of location. Thus, measures of central tendency refer to all those methods of statistical analysis by which averages of the statistical series are worked out.

Definition

According to **Croxton** and **Cowden,** “An average is a single value within the range of the data that is used to represent all of the values in the series. Since an average is somewhere within the range of data**,** it is sometimes called measure of central value. **”**

According to **Clark,** “An average is a figure that represents the whole group.”

2. PURPOSE AND FUNCTIONS OF AVERAGES

Study of averages is of central significance in statistical methods. That is why **Bowley** defines statistics as, “A Science of Averages”.

What is the Basic Purpose of Finding an Average Value of a Series?

It is to identify such a value that represents characteristics of all the items in the series.

According to **Moroney,** “The purpose of an average is brief and simple representation of a group of individual values so that the brain may quickly grasp the general basis of the units of the group. ’’ Some of the main purposes and functions of averages are as under:

(1) Brief Description: The main purpose of an average is to present a brief description of the principal features of the raw data. As a result, data are easily understood.

(2) Comparison: Averages help in making comparison of different sets of data. For example, a comparison of the per capita income of India and USA shows that per capita income of India is much less than the per capita income of USA. Accordingly, it is concluded that India is a poor country.

(3) **Formulation of Policies:** Averages help in formulation of policies. For example, in India the per capita income is Rs. 82,269 per annum which is much less than many countries in the world. Accordingly, it becomes clear for the government to focus on such economic policies as are likely to increase per capita income.

(4) Statistical Analysis: Averages constitute the basis of statistical analysis. For example, if one knows the average marks secured by the students of a class in their different subjects, one can easily analyse the subjects in which the students are weak.

(5) One Value for All: Averages represent the universe or the mass of statistical data. One value represents all values of the series. Accordingly, conclusion can be drawn in respect of the universe as a whole.

3. ESSENTIALS OF A GOOD AVERAGE

A good and satisfactory average should have the following features:

(1) Clear and Stable Definition: A good and a satisfactory average should be clear and stable in definition.

(2) Representative: An average value should be representative of the entire mass of data. It should be based on all the observations of the series.

(3) Simplicity: Simplicity is another essential feature of a good average. It must be so simple that it is easily worked out.

(4) Certainty: A good average must be certain in character. Only then an average value can be used as the basis of statistical analysis.

(5) **Absolute Number:** A good average should be an absolute number. A percentage or a relative value does not serve as a good average.

(6) **Least Effect of a Change in the Sample:** An average of a series should be least affected by a change in the sample on which the average is based.

(7) **Algebraic Treatment:** A good average should be capable of further mathematical or algebraic treatment.

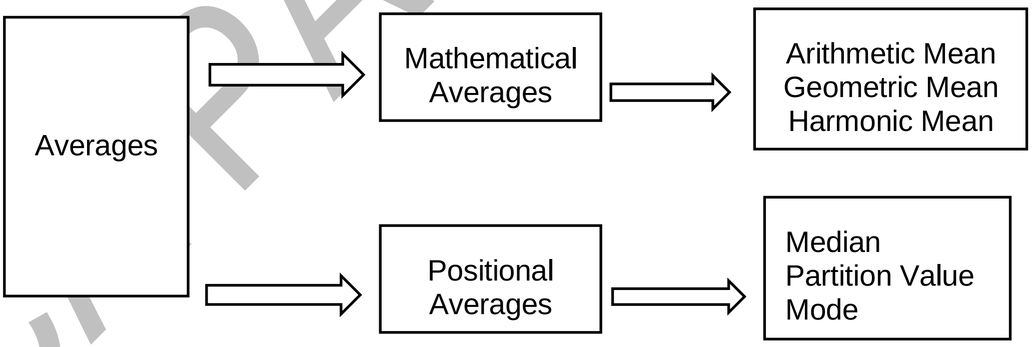
4. TYPES OF STATISTICAL AVERAGES

Averages are broadly classified into two categories:

(1) Mathematical' Averages, and

(2) Positional Averages.

Following chart reveals their further categorisation:



In the present chapter we discuss arithmetic mean.

5. ARITHMETIC MEAN

Arithmetic Mean is a simple average of all items in a series. It is the simplest measure of central tendencies. The arithmetic mean of a series is simply called ‘Mean’. If. for example, Ram plays 5 matches, Shyam 6, Mohan 7, Kishan 8 and Ravi 9 matches a week, the average number of matches played by Ram, Shyam, Mohan, Kishan and Ravi would be determined as under:

Number of Matches = 5+ 6 + 7 + 8 + 9 = 35

Number of Boys = 5

$$\text{Mean} = \frac{\text{Total Value of the Items}}{\text{Number of Items}} = \frac{35}{5} = 7$$

What is the basic difference between Simple Arithmetic Mean and Weighted Arithmetic Mean?

In simple arithmetic mean, all items of the series are taken as of equal importance. In the weighted average, on the other hand, different items are taken as of different importance; accordingly, weights are accorded to different items depending on their relative importance.

Definition

Arithmetic Mean or Mean is the number which is obtained by adding the values of all the items of a series and dividing the total by the number of items.

In the words of **H. Secrist,** “The Arithmetic Mean is the amount secured by dividing the sum of value of the items in a series by their numbers. ”

FORMULA

Arithmetic mean is generally written as X. It may be expressed in the form of following formula:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots \ldots + X_n}{N} = \frac{\Sigma X}{N}$$

Here, $X_{1;}$ X_2, X_3, X_n are the values of different items in the series. Thus, X, = matches played by Ram, 5; X_2= matches played by Shyam, 6 and X_3= matches played by Mohan, 7, etc.

N = Total number of items (in the above example, it is 5 comprising of five boys, Ram, Shyam, Mohan, Kishan and Ravi).

Σ is a sign called Sigma. It refers to the sum total of the values of different items in the series.

Types of Arithmetic Mean

Arithmetic mean is of two types:

(1) Simple Arithmetic Mean: In it, all items of a series are given equal importance.

(2) **Weighted Arithmetic Mean**: In it, different items of a series are accorded different weights in accordance with their relative importance.

Methods of Calculating Simple Arithmetic Mean

We know, there are three types of statistical series:

(1) Individual Series

(2) Discrete Series

(3) Frequency Distribution.

Arithmetic mean may be calculated with respect to these series using different methods.

Calculation of Simple Arithmetic Mean in Case of Individual Series

In the case of individual series, arithmetic mean may be calculated by two methods:

(1) Direct Method

(2) Short-cut Method.

(1) Direct Method

Following steps are involved in this method:

(i) Add up values of all the items of a series (∑X);

(ii) Find out total number of items in the series (N); and

(iii) Divide the total of value of all the items (∑X) with the number of items (N). The resultant value would be the arithmetic mean. Thus,

FORMULA

$$\overline{X} = \frac{\Sigma X}{N} \text{ OR } X = \frac{\text{TOTAL VALUE OF THE ITEMS}}{\text{NUMBER OF ITEMS}}$$

Illustration.

Pocket allowance of 10 students is Rs. 15, 20, 30, 22, 25, 18, 40, 50, 55 and 65. Find out the average pocket allowance.

Solution:

Pocket Allowance (Rs.) (X)
15
20
30

22
25
18
40
50
55
65
∑X = 340

$$\bar{X} = \frac{\Sigma X}{N} = \frac{X_1 + X_2 + \cdots + X_{10}}{10} = \frac{340}{10} = 34$$

Average pocket allowance of the 10 students = Rs. 34.

(2) Short-cut Method

This method is used when the size of items is very large.

The use of short-cut method involves the following steps:

(i) Before finding an actual average, some value in the series is taken as ‘assumed average’. Assumed average should preferably be the middle item of the series. It facilitates the calculation of deviation from the assumed average.

(ii) Assumed average is generally taken by dividing by 2, the total of maximum and minimum values of the items. It is always a complete number. In statistics, assumed average is often denoted as A’.

(iii) Deviations of different values from the assumed average are found and each deviation is written against the concerned value in the series. Thus, d (deviation) = X - A

Where, X is the actual value in the series and A is the assumed average. If value of the item (X) is less than the assumed average (A), then ‘d’ or X - A would be negative. Thus, in the earlier illustration, if A is 40, ‘d’ corresponding to 1st value in the series (15) would be, 15 - 40 = (-) 25. Likewise corresponding to 8th value in the series (50) it would be, 50-40 = (+) 10.

While noting down the deviation against a particular item, sign of the deviation (+) or (-) must also be specified. Thus, ‘d’ of the first value would be written as, - 25 and of the 8th value as, +10.

(iv) Find the sum/total of all deviations. Add up positive deviations and negative deviations separately; and then find out the difference. If the sum of negative deviations is more than that of positive deviations then net sum of all the deviations will be negative and vice versa.

(v) Divide the net sum of the deviations by the number of items in the series. If the dividend is positive (+) then it gets added to A, the assumed average. And, if the dividend is negative (-), then it gets subtracted from A, the assumed average. The value, thus, obtained would be the actual average of the series.

FORMULA

$$\bar{X} = A + \frac{\sum D}{N}$$

(Here, X = Arithmetic mean; A = Assumed average; ∑d = Net sum of the deviations of the different values from the assumed average; N = Number of items in the series.)

Illustration.

Following is the pocket allowance of 10 students. Find out arithmetic mean using Short-cut Method.

Pocket Allowance (Rs.)	15	20	30	22	25	18	40	50	55	65

Solution:

(Assumed Average, A = 40)

Number of Students	Pocket Allowance (Rs.) (X)	Deviation from the Assumed Average (d = X - A), (A = 40)	
1	15	15-40 = -25	
2	20	20-40 = -20	
3	30	30-40 = - 10	
4	22	22-40 = - 18	-110
5	25	25-40 = - 15	
6	18	18-40 = -22	
7	40(A)	40 - 40 = 0	
8	50	50-40 = +10	
9	55	55-40 = +15	+50
10	65	65 - 40 = +25	

N = 10		$\Sigma d = -110 + 50 = -60$

The sum of “+’ deviations = + 50

The sum of deviations =-110

The net sum of deviations, ∑d= - 110 + 50 = - 60

Dividing the aggregate of deviations (∑d) by the number of

items (N), $\frac{\Sigma d}{N} = \frac{-60}{10} = -6$

Substituting this value of $\frac{\Sigma d}{N}$ in the following formula:

$$X = A + \frac{\Sigma D}{N}$$

We have,

X = 40 + (-) 6 = 40-6 = 34

Arithmetic Mean = Rs. 34.

Calculation of Simple Arithmetic Mean in Case of Discrete Series or Frequency Array

Individual series do not have frequencies of the items. These series show only values of different items.

In discrete series, there are frequencies corresponding to different items in the series. There are three methods of calculating mean of the discrete series.

That discrete series (also called frequency array) do not have class intervals. An item in the series does not assume any range of values but each item has corresponding frequency.

(1) Direct Method

(2) Short-cut Method

(3) Step-deviation Method.

(1) Direct Method

Direct method of calculating mean of the discrete series involves the following steps:

(i) Values of the various items in the series are indicated by X, and their frequencies by ‘f’.

(ii) Each item is multiplied by its frequency to get ‘fX’. These multiples are added to get ΣFX. That is,

$\Sigma FX = f_1X_1 + f_2X_2 + \ldots + f_nX_n$

(iii) Frequencies are added up to get ∑f. That is,

$\sum f = f, + f_2 + f_3 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + f_n$ '

(iv) ΣfX is divided by ∑f to obtain the mean, X.

FORMULA

$$\overline{X} = \frac{\Sigma FX}{\Sigma F}$$

Illustration.

Following is the weekly wage earnings of 19 workers:

Wages (Rs.)	10	20	30	40	50
Number of Workers	4	5	3	2	5

Calculate arithmetic mean using Direct Method.

Solution:

Wages (Rs.) (X)	**Number of Workers or Frequency (f)**	**Multiple of the Value of X and Frequency (FX)**
10	4	4 x 10 = 40
20	5	5 x 20 = 100
30	3	3 x 30 = 90
40	2	2 x 40 = 80
50	5	5 x 50 = 250
	∑f = 19	ΣfX = 560

$$\overline{X} = \frac{\Sigma FX}{\Sigma F} = \frac{560}{19} = 29.47$$

Mean wage earnings of 19 workers = Rs. 29.47.

(2) Short-cut Method

Short-cut method of estimating mean of the discrete frequency series uses the following formula:

FORMULA

$$\overline{X} = A + \frac{\Sigma fd}{\Sigma f}$$

Steps

(i) Before calculating the actual average of the series, some value which lies in the middle of the series is taken as assumed average. This may be indicated by ‘A’. In the earlier example, assumed average may be taken as 50.

(ii) Deviation of each value of different items in the series is calculated from the assumed average. The deviation is noted against the concerned item of the series. The deviation may be positive (+) or negative (-). Accordingly, the,(4) or (-) sign is to be noted against each deviation. If the deviation is zero, no sign needs to be specified.

(iii) Find the multiple of ‘d’ and its corresponding ‘f’, that is, ‘fd’. Add up, separately, the positive (+) and negative (-) values of all ‘fd’. Find out the difference between the two to get ‘Σfd.

(iv) Add up all frequencies, to get ∑f.

(v) Divide ∑fd by ∑f that is, $\frac{\Sigma fd}{\Sigma f}$

If the value of $\frac{\Sigma fd}{\Sigma f}$ is positive, it gets added to the assumed average. If this value is negative, it gets subtracted from the assumed average. The result would be the actual average of the series.

Illustration.

Following are the wages of 19 workers:

Wages (Rs.)	10	20	30	40	50
Number of Workers	4	5	3	2	5

Calculate arithmetic mean, using Short-cut Method.

Solution:

(Assumed **Average,** A = 30)

Wages (Rs.) (X)	**Number of Workers or Frequency (f)**	**Deviation (d = X - A) (A = 30)**	**Multiple of Deviation and Frequency (fd)**
10	4	10-30 = - 20	4 x (-20)= - 80 -130
20	5	20-30 = - 10	5 × (-10) = -50
30	3	30 - 30 = 0	3 × 0 =0
40	2	40-30 = 10	2 x 10 =20
50	5	50 - 30 = 20	5 × 20 = 100 +120
	∑f = 19		∑fd = -130 + 120 = -10

$$\overline{X} = A + \frac{\Sigma fd}{\Sigma f} = 30 + \frac{-10}{19} = 30 - \frac{10}{19}$$

= 30 - 0.53 = 29.47

Arithmetic Mean = Rs. 29.47.

(3) Step-deviation Method

This method is a variant of short-cut method. It is adopted when deviations from the assumed mean have some common factor. In the above example, all the deviations (d) can be divided by 10 which is a common factor in this case. The common factor may be indicated by ‘C’. The deviation (d) when divided by the common factor C is called step-deviation. This may be indicated by d’

Steps

(i) Step deviation d' is obtained by dividing the deviation (of the actual value from the assumed average) by the common factor.

$$d' = \frac{X - A}{C} = \frac{d}{C}$$

(Here, d' = Step deviation; C = Common factor; d = Deviation; X = Value of the item; A = Assumed average.)

(ii) Each step deviation is multiplied with its frequency to get fd'. Sum total of the fd' is obtained to get ∑fd'.

(iii) ∑fd' is divided by ∑f, and then multiplied by the common factor ‘C’. The resultant value is added to A to get the actual average of the series. Thus,

FORMULA

$$\overline{X} = A + \frac{\Sigma fd'}{\Sigma f} \times C$$

Illustration.

Wage rate of 19 workers is given below;

Wages (Rs.)	10	20	30	40	50
Number of Workers	4	5	3	2	5

Calculate arithmetic mean using ‘Step-deviation Method.’

Solution:

(Assumed Average, A = 30)

Wages (Rs.) (X)	Number of Workers or Frequency (f)	Deviation (d = X - A) (A = 30)	Step-deviation (C = 10)	Multiple of Step-deviation and Frequency (fd')
10	4	10-30 = -20	$\frac{-20}{10} = -2$	4 x (-2) = - 8
20	5	20-30 = - 10	$\frac{-10}{10} = -1$	5 x (-1) = - 5
30	3	30 - 30 = 0	$\frac{0}{10} = 0$	3×0 = 0
40	2	40-30 = 10	$\frac{10}{10} = 1$	2x1=2
50	5	50 - 30 = 20	$\frac{20}{10} = 2$	5 x 2 = 10
	∑f= 19			∑fd' = -1

$$\bar{X} = A + \frac{\Sigma fd'}{\Sigma f} \times C$$

A = 30, C = 10 and $\frac{\Sigma fd'}{\Sigma f} = \frac{-1}{19} = -0.053$

Putting these values in the formula;

$\bar{X} = 30 + \frac{-1}{19} \times 10$= 30 - 0.053X10

= 30-0.53 = 29.47

Arithmetic Mean = Rs. 29.47.

Calculation of Simple Arithmetic Mean in Case of Frequency Distribution

in case of a frequency distribution, items are classified into different class intervals like 5-10, 10-15, etc. Each class interval has its frequency. There are three methods of calculating mean in frequency distribution:

(1) Direct Method

(2) Short-cut Method

(3) Step-deviation Method.

(1) Direct Method

This method involves the following steps:

(i) The mid-values of the class intervals are calculated. These may be indicated by ‘m To find the mid-value of class intervals, the lower and upper limits of that class are added and then divided by ‘2’. Thus, mid-value of the class 10-20 would be determined as:

$$\text{m} = \frac{l_1 + l_2}{2} = \frac{10 + 20}{2} = \frac{30}{2} = 15$$

(Here, m = Mid-value; $l_↓$ = Lower limit of the class; $/_2$ = Upper limit of the class.)

(ii) Mid-values are multiplied by their corresponding frequencies. The multiples ‘fin’ are added up to get ∑fm.

(iii) ∑fm is divided by ∑f. The resultant value would be the mean value.

FORMULA

$$\bar{X} = \frac{\sum fm}{\sum f}$$

Illustration.

The following table shows marks in English secured by students of Class X in your school in their examination. Calculate mean marks using Direct Method.

Marks	0-10	10-20	20-30	30-40	40-50
Number of Students	20	24	40	36	20

Solution:

Marks	**Mid-value** $(m = \frac{l_1 + l_2}{2})$	**Number of Students or Frequency** **(f)**	**Multiple of Mid-value and Frequency (fm)**
0-10	$\frac{0 + 10}{2} = 5$	20	20 x 5 = 100
10-20	$\frac{10 + 20}{2} = 15$	24	24 × 15 = 360
20-30	$\frac{20 + 30}{2} = 25$	40	40 × 25 = 1,000
30-40	$\frac{30 + 40}{2} = 35$	36	36 × 35 = 1,260

40-50	$\frac{40+50}{2} = 45$	20	20 × 45 = 900
		∑f= 140	∑fm = 3,620

$$\bar{X} = \frac{\Sigma fm}{\Sigma f} = \frac{3{,}620}{140} = 25.86$$

Mean Marks = 25.86.

(2) Short-cut Method

This method involves the following steps:

(i) Mid-values of the classes are determined, and indicated by

(ii) Deviations of the mid-values from the assumed average (A) are determined and indicated by ‘d’. If negative, these deviations are written as -d, and if positive, these are written as -I-d.

(iii) Deviations (d) are multiplied by the frequencies (f) to get ‘fd’. These are then added up to get ∑fd.

(iv) ∑fd is divided by ∑f. The dividend is added to the assumed average. The resultant value would be the mean value.

FORMULA

$$\bar{X} = A + \frac{\sum fd}{\sum f}$$

Illustration.

The following table shows marks secured by the students of a class in an examination in English:

Marks	0-10	10-20	20-30	30-40	40-50
Number of Students	20	24	40	36	20

Calculate mean marks using Short-cut Method.

Solution:

(Assumed Average, A = 25)

Marks **(X)**	**Mid-value** $(m = \frac{l_1 + l_2}{2})$	**Number of Students or Frequency** **(f)**	**Deviation (d = m - A)** **(A = 25)**	**Multiple of Deviation and Frequency** **(fd)**

0-10	$\frac{0+10}{2}=5$	20	5 - 25 = -20	20 × - 20 = - 400
10-20	$\frac{10+20}{2}=15$	24	15 - 25 = -10	24 × -10 = -240
20-30	$\frac{20+30}{2}=25$	40	25 - 25 = 0	40 × 0 = 0
30-40	$\frac{30+40}{2}=35$	36	35-25 = +10	36 × +10 = + 360
40-50	$\frac{40+50}{2}=45$	20	45-25 = +20	20 × + 20 = + 400
		∑f= 140		∑fd = 120

$$\frac{\Sigma fd}{\Sigma f}=\frac{120}{140}=0.86$$

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

$$= 25 + 0.86$$

$$= 25.86$$

Mean Marks = 25.86.

(3) Step-deviation Method

Step-deviation method is a very useful method of calculating mean value in the case of frequency distribution. It involves the following steps:

(i) Find out mid-values of the class intervals, indicated by ‘m\

(ii) Find out deviations of the mid-values from some assumed average. That is,

d = m - A

(iii) Find out step-deviations by dividing the deviations with some Common Factor (C) that is, $d' = \frac{d}{C}$

(iv) Multiply step-deviations (d') with the corresponding frequencies (f). Add up all the multiples to get ∑fd'.

(v) Divide ∑fd' by ∑f and then multiply it by C. The resultant value is added to A to get the mean value. Thus,

FORMULA

$$\overline{X} = A + \frac{\Sigma fd'}{\Sigma f} \times C$$

Illustration.

The following table shows marks obtained by the students of a class in their test in English:

Marks	0-10	10-20	20-30	30-40	40-50
Number of Students	20	24	40	36	20

Calculate arithmetic mean using Step-deviation Method.

Solution:

(Assumed Average, A = 25)

Marks	**Mid-value** $\left(m = \frac{l_1 + l_2}{2}\right)$	**Number of Students or Frequency (f)**	**Deviation (d = m - A) (A = 25)**	**Step-deviation** $\left(d' = \frac{m-A}{c}\right)$ **(C = 10)**	**Multiple of Step- deviation and Frequency (fd)**
0-10	5	20	-20	-2	-40 -64
10-20	15	24	-10	-1	-24
20-30	25	40	0	0	0
30-40	35	36	+ 10	+ 1	+36 76
40-50	45	20	+20	+ 2	+ 40
		$\sum$f= 140			$\Sigma fd' = 12$

$$\frac{\Sigma fd'}{\Sigma f} = \frac{12}{140} = 0.086$$

$$\text{Mean, } \overline{X} = A + \frac{\Sigma fd'}{\Sigma f} \times C$$

= 25 + 0.086 × 10

= 25 + 0.86 = 25.86

Arithmetic Mean = 25.86 marks.

Calculation of Arithmetic Mean in Case of Cumulative Frequency Distribution

Here are two illustrations to facilitate your understanding.

Illustration 1.

Marks in Statistics of the students of Class XI are given below. Find out arithmetic mean.

Marks	Number of Students
Less than 10	5
Less than 20	17
Less than 30	31
Less than 40	41
Less than 50	% 49

Solution:

A Cumulative Frequency Distribution should first be converted into a Simple Frequency Distribution, as under:

Conversion of a Cumulative Frequency Distribution into a Simple Frequency Distribution

Marks	Number of Students
0-10	5
10-20	17- 5 = 12
20-30	31 - 17 = 14
30-40	41-31 = 10
40-50	49-41 = 8

Now, mean value of the data is obtained using Direct Method as under:

Calculation of Mean

Marks (X)	Mid-value $(m = \frac{l_1 + l_2}{2})$	Number of Students or Frequency (f)	Multiple of Mid-value and Frequency (fm)
0-10	5	5	25
10-20	15	12	180
20-30	25	14	350
30-40	35	10	350
40-50	45	8	360

		$\sum f = 49$	$\sum fm = 1,265$

$$\overline{X} = \frac{\Sigma fm}{\Sigma f} = \frac{1,265}{49} = 25.82$$

Arithmetic Mean = 25.82 marks.

Illustration. 2

The following table shows marks in economics of the students of a class. Calculate arithmetic mean.

Marks	Number of Students
More than 0	30
More than 2	28
More than 4	24
More than 6	18
More than 8	10

Solution:

Converting Cumulative Frequency Distribution into a Simple Frequency Distribution, we get the following:

Marks	Number of Students
0-2	30 - 28 = 2
2-4	28 - 24 = 4
4-6	24- 18 = 6
6-8	18-10 = 8
8-10	10- 0 = 10

Arithmetic mean of this continuous series is estimated below, using Direct Method.

Calculation of Arithmetic Mean using Direct Method

Marks	Mid-value $(m = \frac{l_1 + l_2}{2})$	Number of Students or Frequency (f)	Multiple of Mid-value and Frequency (fm)
0-2	1	2	2

2-4	3	4	12
4-6	5	6	30
6-8	7	8	56
8-10	9	10	90
		∑f = 30	∑fm = 190

$$\bar{X} = \frac{\Sigma fm}{\Sigma f} = \frac{190}{30} = 6.33$$

Arithmetic Mean — 6.33 marks.

Calculation of Arithmetic Mean in a Mid-value Series

Here is an illustration.

Illustration.

Following table gives marks in Statistics of the students of a class. Find out mean marks.

Mid-value	5	10	15	20	25	30	35	40
Number of Students	5	7	9	10	8	6	3	2

Solution:

In this series, mid-values are already given. The calculation of arithmetic mean involves the same procedure as in the case of exclusive series.

Calculation of Arithmetic Mean

Mid-value **(m)**	**Number of Students or Frequency** **(f)**	**Multiple of Mid-value and Frequency (fm)**
5	5	25
10	7	70
15	9	135
20	10	200
25	8	200
30	6	180
35	3	105

40	2	80
	∑f = 50	∑fm = 995

$$\bar{X} = \frac{\Sigma fm}{\Sigma f} = \frac{995}{50} = 19.9$$

Mean Marks =19.

Calculation of Arithmetic Mean in Case of Inclusive Series

Illustration.

The following table shows monthly pocket expenses of the students of a class. Find out the average pocket expenses.

Pocket Expenses (Rs.)	20-29	30-39	40-49	50-59	60-69
Number of Students	10	8	6	4	2

Solution:

Calculation of arithmetic mean of inclusive series is the same as of exclusive series.

Pocket Expenses (Rs.)	**Mid-value** $(m = \frac{l_1 + l_2}{2})$	**Number of Students or Frequency (f)**	**Deviation (d = m - A) (A - 44.5)**	**Step deviation** $\left(d' = \frac{d}{c}\right)$ **(C = 10)**	**Multiple of Step-deviation and Frequency (fd)**
20-29	24.5	10	-20	_ 2	-20
30-39	34.5	8	- 10	-1	-8
40-49	44.5	6	0	0	0
50-59	54.5	4	+ 10	+ 1	+ 4
60-69	64.5	2	+20	+ 2	+ 4
		∑f = 30			∑fd' = - 20

$$\bar{X} = A + \frac{\Sigma fd'}{\Sigma f} \times C$$

$$= 44.5 + \frac{-20}{30} \times 10$$

$$= 44.5 - \frac{20}{30}$$

= 44.5-6.67 = 37.83

Average Pocket Expenses = Rs. 37.83.

Calculation of 'Corrected' Arithmetic Mean

Sometimes, in the calculation of Arithmetic Mean, some items or values are wrongly written. Accordingly, the mean value goes wrong. But such mean value can be corrected, as illustrated below: $\bar{X} = \frac{\Sigma X(\text{Wrong}) + (\text{ Correct value }) - (\text{ Incorrect value })}{N}$

Illustration.

Mean marks obtained by 100 students are estimated to be 40. Later on it is found that one value was read as 83 instead of 53.

Find out the “corrected” mean.

Solution:

$$\bar{X} = \frac{\Sigma X}{N}$$

Or, $\Sigma X = \bar{X}N$

Given: $\bar{X}$ = 40; N = 100 $\sum X$

$$40 = \frac{\Sigma X}{100}$$

$\sum X$(wrong) = 40×100 = 4,000

Correct value = 53

Incorrect value = 83

Correct

$$\bar{X} = \frac{\sum X(\text{WRONG}) + (\text{ CORRECT VALUE }) - (\text{ INCORRECT VALUE })}{N}$$

$$= \frac{4{,}000 + 53 - 83}{100}$$

$$= \frac{3{,}970}{100} = 39.70$$

Thus, Corrected Mean = 39.70.

Finding the Missing Value

If mean of the series is known, but one value is missing, the missing value may be found using the following formula:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{N}$$

Illustration.

Suppose mean of a series of 5 items is 30, Four values are, 10, 15, 30 and 35 respectively. Find the missing (5th) value of the series.

Solution.

Assume 5th value as X_5.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{N}$$

Given: $X_1 = 10$, $X_2 = 15$, $X_3 = 30$, $X_4 = 35$, $X_5 = ?$

$\overline{X} = 30$, $N = 5$

$$30 = \frac{10 + 15 + 30 + 35 + x_5}{5} = \frac{90 + x_5}{5}$$

∴ $30 \times 5 - 90 + X_5$

Or, $150 = 90 + X_5$

∴ $X_5 = 150 - 90 = 60$

Thus, Value of the 5th item = 60.

WEIGHTED ARITHMETIC MEAN

We have so far discussed Simple Arithmetic Mean. In simple arithmetic mean, all items of a series are taken as of equal significance. But sometimes we may give greater significance to some items and less to others. As household may, for example, give more significance to food, less to clothes and still less to entertainment. When different items of a series are weighed according to their relative importance, the average of such series is called Weighted Arithmetic Average. A Weighted Arithmetic Average is, thus, the mean of weighted items.

Calculation of Weighted Mean

Calculation of weighted mean involves the following steps:

(i) Different items are weighed according to their significance. Weights are indicated by ‘W’.

(ii) Items (X) are multiplied by their corresponding weights (W) and added up to get ∑WX.

(iii) ∑WX is divided by the sum total of weights, i.e., ∑W, to get the mean value, that is,

FORMULA

$$\bar{X}_w = \frac{\Sigma WX}{\Sigma W}$$

Here, $\bar{X}_w$ indicates weighted average.

Illustration.

Calculate weighted mean of the following data:

Marks (X)	81	76	74	58	70	73
Weight (W)	9	3	6	7	3	7

Solution:

Calculation of Weighted Mean

Marks (X)	**Weight (W)**	**wx**
81	2	162
76	3	228
74	6	444
58	7	406
70	3	210
73	7	511
	$\sum W = 28$	$\sum WX = 1{,}961$

Weighted Mean,

$$\bar{X}_w = \frac{\Sigma WX}{\Sigma W} = \frac{1{,}96l}{28} = 70.04$$

Weighted Mean = 70.04 marks.

COMBINED ARITHMETIC MEAN

Given the mean values of two or more parts of a series and the number of items in each part, one can get Combined Arithmetic Mean, or mean of the series as a whole. The following formula is used for the estimation of combined arithmetic mean:

FORMULA

$$\bar{X}_{1,2} = \frac{\bar{X}_1 N_1 + \bar{X}_2 N_2}{N_1 + N_2}$$

Here,

$\bar{X}_{1,2}$ = Combined arithmetic mean of parts 1 and 2 of a series;

$\bar{X}_1$ = Arithmetic mean of part 1 of the series;

$\bar{X}_2$ = Arithmetic mean of part 2 of the series;

N_1 = Number of items in part 1 of the series;

N_2 = Number of items in part 2 of the series.

Likewise, when there are more than 2 parts ofa series, the following formula is used to work out Combined Arithmetic Mean. *

FORMULA

$$\bar{X}_{1,2,3,\ldots,n} = \frac{\bar{X}_1N_1 + \bar{X}_2N_2 + \cdots + \bar{X}_nN_n}{N_1 + N_2 + \cdots + N_n}$$

Illustration.

60 students of Section A of Class XI, obtained 40 mean marks in Statistics, 40 students of Section B obtained 35 mean marks in Statistics. Find out mean marks in Statistics for Class XI as a whole.

Solution:

Given:

$N_1 = 60$, $\bar{X}_1 = 40$, $N_2 = 40$, $\bar{X}_2$ - 35

We know, $\bar{X}_{12} = \frac{\bar{X}_1N_1+\bar{X}_2N_2}{N_1+N_2}$

$$\bar{X}_{12} = \frac{40 \times 60 + 35 \times 40}{60 + 40}$$

$$\bar{X}_{12} = \frac{2{,}400 + 1{,}400}{100} = \frac{3{,}800}{100} = 38$$

Thus, Combined Arithmetic Mean = 38.

Addition, Subtraction, Multiplication and Division of Values

If some specific value is added to or subtracted from different items in a series, the mean value of the items would increase or decrease by the same specific value respectively. Likewise, if different items of a series are multiplied or divided by any specific value, the mean value of the items would be the multiple or divided by the same specific value. The following example illustrates this point:

X	(X+2)	(X-2)	(X × 2)	(X + 2)
10	12	8	20	5
20	22	18	40	10
30	32	28	60	15
40	42	38	80	20
50	52	48	100	25

ΣX = 150	ΣX = 160	ΣX = 140	ΣX = 300	ΣX = 75
N = 5	N = 5	N = 5	N = 5	N = 5
X = 30	X = 32	X = <28	X = 60	X = 15

The example shows that: (i) if 2 is added to different items, the resultant average of the series also increases by 2 (30 + 2 = 32); (ii) if 2 is subtracted from the different items, the resultant average also decreases by 2 (30 - 2 = 28); (hi) if the items are multiplied by 2, the resultant average also increases by the multiple of 2 (30 × 2 = 60); and (iv) if different items are divided by 2 the resultant average will also be divided by 2 (30 => 2 = 15).

Location of Missing Item or Frequency

A missing item (referring to any value of a variable X) or a missing frequency (f) of a series can be located using the following equations:

$\bar{X} = \frac{\Sigma X}{N}$ $\qquad$ $\bar{X} = \frac{\Sigma fX}{\Sigma f}$ $\quad (N = \Sigma f)$

$\Sigma X = \bar{X}N$ $\qquad$ $\Sigma fX = \bar{X}\Sigma f$

$N = \frac{\Sigma X}{\bar{X}}$ $\qquad$ $\Sigma f = \frac{\Sigma fX}{\bar{X}}$

Illustration 1.

In the following frequency distribution, the frequency of the class interval (40-50) is not known. Find it, if the arithmetic mean of the distribution is 52.

Wages (Rs.)	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of Workers	5	3	4		2	6	13

Solution:

Wages (Rs.)	**Mid-value (m)**	**Number of Workers or Frequency (f)**	**Multiple of Mid-value and Frequency (fm)**
10-20	15	5	75
20-30	25	3	75
30-40	35	4	140
40-50	45	f	45 f
50-60	55	9	no
60-70	65	6	390
70-80	75	13	975

		$\sum f = 33 + f$	$\sum fm = 1{,}765 + 45f$

$$\bar{X} = \frac{\sum fm}{\Sigma f}$$

$$52 = \frac{1{,}765 + 45f}{33 + f}$$

Or, 52 (33 + f) = 1,765 + 45f

1,716 + 52f = 1,765 + 45f

52f - 45f = 1,765- 1,716

=> 7f = 49

=> f = 7

Missing Frequency = 7.

Illustration 2.

If the arithmetic mean of the following series is 115.86; find the missing value.

Wages (Rs.)	110	112	113	117	P	125	128	130
Number of Workers	25	17	13	15	14	8	6	2

Solution:

Let the missing value be X.

Wages (Rs.) (X)	**Number of Workers or Frequency (f)**	**Multiple of the Value of X and Frequency (fX)**
110	25	2,750
112	17	1,904
113	13	1,469
117	15	1,755
X	14	14X
125	8	1,000
128	6	768
130	2	260
	$\sum f = 100$	$\Sigma fX = 9{,}906+14X$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f}$$

$$115.86 = \frac{9{,}906 + 14X}{100}$$

$$115.86 \times 100 = 9{,}906 + 14X$$

$$14X = 11{,}586\text{-}9{,}906$$

$$14X = 1{,}680$$

$$X = \frac{1{,}680}{14} = 120$$

Missing Value = 120.

A Notable Property of Arithmetic Mean

A notable property of arithmetic mean is that the sum of deviations of different items of a series, when deviations are taken from arithmetic mean, is always zero. This important notable property of arithmetic mean can be proved by the following example. The sum of the deviations from its arithmetic mean would be zero. This is worked out as under:

Illustration.

Show that in the series of marks 10, 20, 30, 40, 50 obtained by five students the sum of deviations from arithmetic mean is zero.

Solution:

Marks **(X)**	**$\bar{X}$ = 30** **Deviations from Arithmetic Mean (X - $\bar{X}$)**
10	10-30 =-20
20	20-30 = - 10
30	30 - 30 = 0
40	40-30 = + 10
50	50 - 30 = + 20

ΣX = 150
N = 5
X = 30
	2(X- $\bar{X}$) = 0

Thus, the sum of the deviations from arithmetic mean Σ(X — X) = zero.	

Properties of Arithmetic Mean (AM)

(i) The sum of deviations of the items from AM is always zero.

(ii) The sum of squared deviations of the items from AM is minimum.

(iii) If each item of a series is increased, decreased, multiplied or divided by some constant then AM also increases, decreases, multiplies or is divided by the same constant.

(iv) The product of the AM and the number of items on which mean is based is equal to the sum of all given items.

(v) if each item of the original series is replaced by the actual mean, then the sum of these substitutions will be equal to the sum of the individual items.

The Principal Merit of Arithmetic Mean

That it is based on all items of the series and is capable of further algebraic treatment.

The Principal Demerit of Arithmetic Mean

End-use of the goods is the principal basis of classifying the goods as intermediate goods and final goods.

Merits and Demerits of Arithmetic Mean

Merits

The following are some of the main merits of arithmetic mean:

(1) Simple: From the viewpoint of calculation and usage, arithmetic mean is the simplest of all the measures of central tendency.

(2) Certainty: Arithmetic mean is a certain value; it has no scope for estimated values.

(3) Based on All Items: Arithmetic mean is based on **all** the items in a series. It is, therefore, a representative value of the different items.

(4) Algebraic Treatment: Arithmetic mean is capable of further algebraic treatment. It is, therefore, extensively used in statistical analysis.

(5) Stability: Arithmetic mean is a stable measure of central tendency. This is because changes in the sample of a series have minimum effect on the arithmetic average.

(6) Comparison: Being stable and certain, arithmetic mean can be easily used for comparisons.

(7) Accuracy Test: Arithmetic mean can be tested for its accuracy as a representative value of the series.

Demerits

Arithmetic mean suffers from following demerits:

(1) Effect of Extreme Value: The main defect of arithmetic mean is that it gets distorted by extreme values of the series. Therefore, it is not always an accurate measure. To illustrate, pocket expenditure of a rich student of Class XI may be Rs. 2,000, while his four friends incur pocket expenditure of Rs. 100, Rs. 80, Rs. 70 and Rs. 50 respectively. The average pocket expenditure of all the five students would be:

$$\bar{x} = \frac{2{,}000 + 100 + 80 + 70 + 50}{5} = \frac{2{,}300}{5} = \vec{7}460$$

Certainly this is not such an accurate mean of the pocket expenditure of five students as 4 out of 5 students incur pocket expenditure just ranging between Rs. 50 to Rs. 100. The mean value of Rs. 460 is largely owing to the fact that there is an extreme value of Rs. 2,000 in the series. Such a mean is certainly not a representative value of all the items in the series.

(2) Mean Value may not figure in the series at all: The mean value may sometimes be that value which does not figure in the series at all.

The average of 2, 3 and 7 is $\frac{2+3+7}{3} = 4$ which is not there in the series. It further erodes its representative character.

(3) Laughable Conclusions: Arithmetic mean sometimes offers laughable conclusions. If there are 50 students in Class XI and 51 students in Class XII, the average strength of these two classes would come to $\frac{50+51}{2} = 50.5$ students, which is indeed very funny because there cannot be half student.

(4) Unsuitability: Arithmetic mean is not a suitable measure in case of percentage or proportionate values.

(5) Misleading Conclusions: Arithmetic mean sometimes offers misleading conclusions. Following example illustrates this point:

Income of Two Individuals A and B for the years 2016-2018

Year	**Income (Rs.)**	
	A	**B**
2016	1,000	3,000
2017	2,000	2,000
2018	3,000	1,000
	6 000	
	$\bar{X} = \frac{6{,}000}{3} = 2{,}000$	$\bar{X} = \frac{6{,}000}{3} = 2{,}000$

In the given Table, average income of both A and B is same suggesting that A and B have been equally rich during the years 2016 to 2018. But a close examination of the series should reveal that while the income of A is increasing over years, that of B is decreasing. Thus, arithmetic mean offers misleading conclusions.

However, despite certain demerits of arithmetic mean as noted above, this is an ideal measure of the central tendency. this is the most widely used measure in practical life. Arithmetic mean is particularly significant in such series of which different items are equally important and therefore, equally weighed. Average output, average cost, average revenue, are some of the well-known concepts in Economics based on arithmetic mean.

Calculation of Arithmetic Mean for Different Series: A Glance

Series	Methods of Calculation of Arithmetic Mean
1. Individual Series	(a) Direct Method: Formula: $\bar{X} = \frac{\Sigma X}{N}$
	(b) Short-cut Method: Formula: $\bar{X} = A + \frac{\Sigma d}{N}$
2. Discrete Series/Frequency Array	(a) Direct Method: Formula: $\bar{X} = \frac{\Sigma fX}{\Sigma f}$
	(b) Short-cut Method: Formula: $\bar{X} = A + \frac{\Sigma fd}{\Sigma f}$
	(c) Step-deviation Method: Formula: $\bar{X} = A + \frac{\Sigma fd'}{\Sigma f} \times C$
3. Frequency Distribution Series	(a) Direct Method: Formula: $\bar{X} = \frac{\Sigma fm}{\Sigma f}$
	(b) Short-cut Method: Formula: $\bar{X} = A + \frac{\Sigma fd}{\Sigma f}$
	(c) Step-deviation Method: Formula: $\bar{X} = A + \frac{\Sigma fd'}{\Sigma f} \times C$

Multiple Choice Questions

Select the correct alternative:

1. Which of the following is a type of mathematical average?

(a) Median

(b)Partition value

(c) Mode

(d) None of these

2. Formula for finding arithmetic mean is:

(a) $\bar{x} = \Sigma x$

(b) $\bar{x} = \frac{\Sigma x}{N}$

(c) $\bar{X} = \Sigma X - N$

(d) $\bar{x} = \frac{N}{\Sigma X}$

3. Arithmetic mean of these items 5, 7, 9, 15, 20 is:

(a) 10

(b) 10.2

(c) 11.2

(d) 12

4. Arithmetic mean of these items: 10,15, X, 20, 30 is 20. Find out the missing item.

(a) 10

(b) 15

(c) 5

(d) 25

5. By which formula is combined arithmetic mean estimated?

(a) $\bar{x}_{12} = \frac{x_1+x_2+\cdots+x_n}{N_1+N_2}$

(b) $\bar{X}_{12} = \frac{\bar{X}_1 N_1+\bar{X}_2 N_2}{N_1+N_2}$

(c) $\bar{X}_{12} = \frac{\bar{X}_1+\bar{X}_2}{N_1+N_2}$

(d) None of these

6. Arithmetic mean of a series is 15 and if 5 is added in all the items of this series, the new arithmetic mean will be:

(a) 5

(b) 20

(c) 18

(d)10

7. What is the formula to find out arithmetic mean through Short-cut Method in individual series?

(a) $\bar{x} = \frac{\Sigma x}{N}$

(b) $\bar{x} = A + \frac{\Sigma d}{N}$

(c) $\bar{x} = \frac{\Sigma X}{N} + A$

(d) $\bar{x} = \frac{\Sigma fX}{\Sigma f}$

8. Which of the following is not a measure of central tendency?

(a) Mean

(b) Mode

(c) Standard deviation

(d) Median

9. Which is not a method to find arithmetic mean?

(a) Direct method

(b) Short-cut method

(c) Step-deviation method

(d) Karl Pearson's method

10. Assumed mean is taken in which method?

(a) Direct method

(b) Step-deviation method

(c) Karl Pearson's method

(d) Spearman's method

11. Sum of deviations of different values from arithmetic mean is always equal to:

(a) zero

(b) one

(c) less than one

(d) more than one

12. Mean of 0.3, 5, 6, 7, 9,12, 0.6 is:

(a) 4.9

(b) 5.7

(c) 5.6

(d) None of these

13. Simple average is sometimes called:

(a) Unweighted average

(b) Weighted average

(c) Relative average

(d) None of these

14. __________ is used when the sum of deviations from the average should be least.

(a) Mean

(b) Mode

(c) Median

(d) None of these

15. Mean should be:

(a) Simple

(b) Based upon all items

(c) Rigidly defined

(d) Ail the above

16. The algebraic sum of deviations of 8,1,6 from the A.M. viz., 5 is:

(a) -1

(b) 0

(c) 1

(d) None of these

17. Average value of given variables is known as:

(a) Median

(b) Mean

(c) Mode

(d) Index Number

18. Measures of central tendency are known as:

(a) Difference

(b) Averages

(c) Both

(d) None of these

19. The A.M. of 1,3, 5, 6, x, 10 is 6. The value of x is:

(a) 10

(b) 11

(c) 12

(d) None of these

20. The values of all items are taken into consideration in the calculation of:

(a) Median

(b) Mode

(c) Mean

(d) None of these

21. Sum of the deviations about mean is:

(a) Zero

(b) Minimum

(c) Maximum

(d) One

22. Measures of central tendency for a given set of observations measures:

(a) The scatterness of the observations

(b) The central location of the observations

(c) Both (a) and (b)

(d) None of these

23. If there are two groups containing 30 and 20 observations and having 50 and 60 as arithmetic means, then the combined arithmetic mean is:

(a) 51

(b) 54

(c) 53

(d) 52

24. The mean of 12 numbers is 24. If 5 is added in every number, the new mean is:

(a) 25

(b) 84

(c) 29

(d) None of these

25. Sum of square of the deviations about mean is:

(a) Maximum

(b) Minimum

(c) Zero

(d) None of these

26. __________ average is obtained by dividing the total of set of observations by their number.

(a) Weighted

(b) Simple

(c) Both (a) and (b)

(d) Neither (a) nor (b)

Answers

Multiple Choice Questions

1. (d)	**2. (b)**	**3. (c)**	**4.(d)**	**5. (b)**
6. (b)	**7. (b)**	**8. (c)**	**9. (d)**	**10. (b)**

11. (a) **12. (b)** **13. (a)** **14. (a)** **15. (d)**

16. (b) **17. (b)** **18. (b)** **19. (b)** **20. (c)**

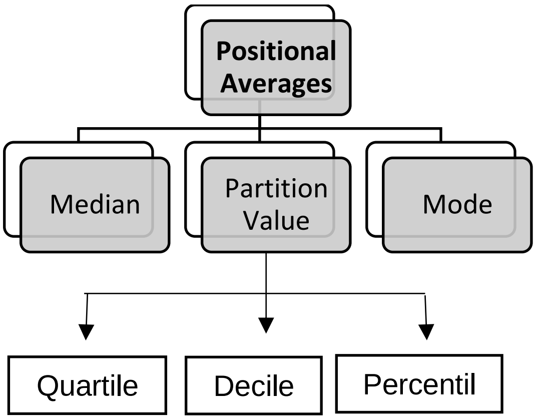
21. (a) **22. (b)** **23. (b)** **24. (c)** **25. (b)**

26. (b)

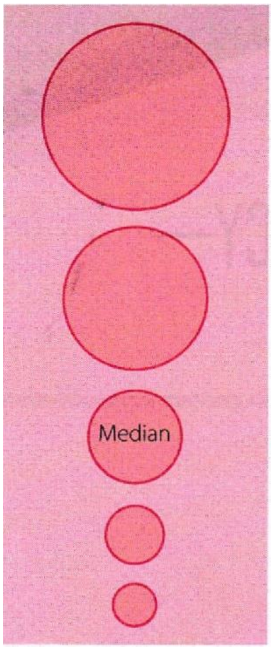
CHAPTER 10

MEASURES OF CENTRAL TENDENCY- MEDIAN AND MODE

In a statistical series, there may be a value which is centrally located or which occurs most frequently in the series. This is known as central value of the series. For example, if five friends watch, 5, 10, 15, 20 and 30 movies respectively, in a month, then 15 is the central value of the series. Take another example, suppose 80 out of 100 customers visiting a shoe shop buy shoes of size 8, then 8 occurs most frequently in a series of 100 observations. Accordingly, 8 will be taken as the central value of the series, though it includes other values as well like, 7,9, 10, etc. In the first example, ‘15’ is the central value and it is positioned somewhere in the middle of the series. Hence, it is called median value. In the second example, 8 happens to occur most frequently in the series (80 times out of 100). Hence, it is called modal value. Median and mode are the most important positional averages. These are called positional averages because their value is worked out on the basis of their position in the statistical series. There are other positional averages as well (besides median and mode) as indicated in the following flow chart.



The present chapter focuses only on median and mode, in accordance with the CBSE guidelines. Only a passing reference is made to the estimation of one partition value, viz, 'Quartile' as its knowledge would help understand the estimation of 'Quartile Deviation' which is one of the various measures of dispersions, discussed in the next chapter.



1. MEDIAN

Median is a centrally located value of a series such that half of the values (or items) of the series are above it and the other half below it. To illustrate if median height of the students of a class is to be determined, all the students may be asked to stand in the increasing or decreasing order of their heights. The student figuring in the middle would be taken as the central item of the series. Height of this student would be taken as median height of the students, representing the whole set of students. A notable point is that the number of values above the central value should be equal to the number of values below it. Mid-value is thus called Median.

Definition

According to **Connor,** "The Median is that value of the variable which divides the group into two equal parts, one part comprising all values greater than the median value and the other part comprising all the values smaller than the median value. "

How to Find the Median Value?

In Statistics, median is indicated by M. In order to locate the median value, all items of a series are arranged in either the ascending order or descending order. In ascending order, higher values follow the lower values. In the descending order, lower values follow the higher values. The median value is located, using the following formula:

FORMULA

$$M = \text{ Size of } \left(\frac{N+1}{2}\right) \text{ th item}$$

(Here, M = Median; N = Number of items.)

Let us learn the application of this formula, taking different types of statistical series.

Calculation of Median for Different Types of Statistical Series

(1) Individual Series and the Median

Calculation of median in individual series involves the following steps:

(i) Arrange all the values of different items of a series in the ascending or descending order.

(ii) Add up the number of items, indicated by N. Find out the median item as $\left(\frac{N+1}{2}\right)$th item. That is, M = Size of $(\frac{N+1}{2})$th item

(Here, M = Median; N = Number of items.)

(iii) If N of series happens to be an odd number, that is, not divisible by 2, it is easier to find out median with the help of the above formula. For example, if the series is 4, 5, 6, 7, 8, then Median will be $\frac{5+1}{2}$ = 3rd item.

4 5 6 7 8

Median

1 2 3 4 5

Median = Size of $(\frac{N+1}{2})$ th =3rd item = 6

(iv) If N of series happens to be an even number, $\frac{N+1}{2}$ will come in fractions. Thus, if the series is 4,10,12, and 18, then the $\frac{N+1}{2}$ would not be a complete number, but it would be like 2.5th value. In such cases, median value would be the arithmetic mean of the two middle values of the series. Thus, median value or the size of 2.5th item would be located as under:

$$\text{Median} = \frac{\text{Size of 2 nd item} + \text{Size of 3 rd item}}{2}$$

$$= \text{Size of 2.5th item}$$

$$= \frac{10+12}{2} = \frac{22}{2} = 11$$

When N of the series is an even number, median is estimated using the following

formula: $M = \frac{\text{Size of } (\frac{N}{2})th \text{ th item} + \text{Size of } (\frac{N}{2}+1) \text{ th item}}{2}$

Illustration.

The following series show marks in economics of 11 students of Class XL Find the median marks.

Marks	17	32	35	33	15	21	41	32	11	10	20

Solution:

Ascending Order		**Descending Order**	
S. No.	Marks	S. No.	Marks
I	10	1	41
2	11	2	35
3	15	3	33
4	17	4	32
5	20	5	32
6 (M)	21	6 (M)	21
7	32	7	20
8	32	8	17
9	33	9	15
10	35	10	11
11	41	11	10
N = 11		N = 11	

$$M = \text{Size of } \left(\frac{N+1}{2}\right) \text{ th item}$$

$$= \text{Size of } \left(\frac{11+1}{2}\right) \text{ th item}$$

$$= \text{Size of 6th item} = 21$$

Median = 21.

(2) Discrete Series or Frequency array and the Median

Calculation of median in case of discrete series or frequency array involves the following steps:

(i) Arrange the data into ascending (or descending) order.

(ii) Convert the simple frequencies of a series into cumulative frequencies,

(iii) Determine the $\left(\frac{N+1}{2}\right)$ th item of the series, N being equal to ∑f.

(iv) Find the median value corresponding to the $\left(\frac{N+1}{2}\right) th$ item

Illustration.

Find the median of the following series:

Size	2	3	4	5	6	7	8	9	10
Frequency	2	3	8	10	12	16	10	8	6

Solution:

Size	Frequency	Cumulative Frequency
2	2	2
3	3	2 + 3 = 5
4	8	5 + 8 = 13
5	10	13 + 10 = 23
6	12	23 + 12 = 35
7	16	35 + 16 = 51 (M)
8	10	51 + 10 = 61
9	8	61 + 8 = 69
10	6	69 + 6 = 75
	N=75	

Median or M= Size of $(\frac{N+1}{2})$ item

= Size of $(\frac{75+1}{2})$ th item

= Size of 38th item

It shows that median value corresponds to the 38th item in the series. This item appears first of all in 51st cumulative frequency of the series. Therefore, median shall be the value corresponding to the 51st cumulative frequency, which is 7.

Median = 7.

(3) Frequency Distribution Series and the Median

Calculation of median in a continuous series involves the

following steps:

(i) The data are arranged in ascending or descending orders of their class interval.

(ii) The frequencies are then converted into cumulative frequencies.

(iii) Median class of the series is identified. It corresponds to that cumulative frequency which includes the 1th item.

(iv) The following formula is applied to determine the actual median value.

FORMULA

$$M = l_1 + \frac{\frac{N}{2} - \text{c.f.}}{\text{f}} \times i$$

(Here, /, = Lower limit of the median class; c.f. = Cumulative frequency of the class preceding the median class; f = Frequency of the median class; i = Size of the median class interval.)

Illustration.

Find out median value of the following distribution:

Wage Rate (Rs.)	0-10	10-20	20-30	30-40	40-50
Number of Workers	22	38	46	35	20

Solution;

Wage Rate (Rs.)	Frequency (f)	Cumulative Frequency
0-10	22	22
10-20	38	60 (c.f.)
(h) 20-30	46(f)	106
30-40	35	141
40-50	20	161
	$\Sigma f = N = 161$	

∴ M = Size of $(\frac{N}{2})$ th item; N = Σf $= 161$

∴ M = Size of $(\frac{161}{2})$ th item

= Size of 80.5th item

80.5th item lies in 106th cumulative frequency. The class interval corresponding to this cumulative frequency is 20-30, which, therefore, is the median class interval. That is, the value of the median must lie within the range of 20-30. The following formula is applied to identify the exact value of the median.

$$M = l_1 + \frac{\frac{N}{2} - \text{c.f.}}{\text{f}} \times i$$

$$= 20 + \frac{\frac{161}{2} - 60}{46} \times 10 = 20 + \frac{80.5 - 60}{46} \times 10$$

$$= 20 + \frac{20.5}{46} \times 10 = 20 + 4.46 = 24.46$$

Median = Rs. 24.46.

Illustration.

The following table gives distribution of marks secured by some students:

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of Students	42	38	120	84	48	36	31

Calculate the median marks secured by the students.

Solution:

Marks	Frequency (f)	Cumulative Frequency
10-20	42	42
20-30	38	80 (c.f.)
(/,) 30-40	120(f)	200
40-50	84	284
50-60	48	332
60-70	36	368
70-80	31	399
	N = 399	

M = Size of $(\frac{N}{2})$ th item

= Size of $(\frac{399}{2})$ th item

= Size of 199.5th item

Hence, median lies in the class 30^0.

$$\text{M} = l_1 + \frac{\frac{\text{N}}{2} - \text{c.f.}}{\text{f}} \times i$$

$$= 30 + \frac{\frac{399}{2} - 80}{120} \times 10$$

$$= 30 + \frac{199.5 - 80}{120} \times 10 = 30 + \frac{119.5}{120} \times 10$$

$$= 30 + 9.96 = 39.96$$

Median Marks = 39.96.

(4) Cumulative Frequency Series and the Median

Illustration.

Calculate median of the following series:

Wage Rate (Rs.) (less than)	10	20	30	40	50	60	70	80
Number of Workers	15	35	60	84	96	127	198	250

Solution:

As a first step the cumulative frequency of ‘less than’ type is converted into a simple frequency distribution as under:

Wage Rate (Rs.)	Cumulative Frequency	Frequency (f)
0-10	15	15
10-20	35	35- 15 = 20
20-30	60	60 - 35 = 25
30-0	8*4	84 - 60 = 24
40-50	96 (c.f.)	96-84 = 12
(/) 50-60	127	127-96 = 31(f)
60-70	198	198- 127 = 71
70-80	250	250- 198 = 52
		∑f = N = 250

M = Size of N/2th item or 250/2 = 125th item which lies in 127th cumulative frequency of the series. The corresponding class interval, 50-60 is, therefore, the median class interval. The actual value of the median is determined using the standard formula, viz.

$$\text{M} = l_1 + \frac{\frac{\text{N}}{2} - \text{c.f.}}{\text{f}} \times i$$

Imputing the values,

$$l_1 = 50, \text{c.f.} \ = 96, \text{f} = 31, i = 10$$

$$\text{M} = 50 + \frac{\frac{250}{2} - 96}{31} \times 10$$

$$= 50 + \frac{125 - 96}{31} \times 10$$

Median Wage Rate = Rs. 59.35.

(5) Inclusive Series and the Median

Illustration.

Calculate median of the following data:

Marks	46-50	41-45	36-40	31-35	26-30	21-25	16-20	11-15
Number of Students	5	11	22	35	26	13	10	7

Solution:

This is an inclusive series given in the descending order. It should first be converted into an exclusive series and placed in the ascending order, as in the following table:

Conversion into Exclusive Series	Frequency (f)	Cumulative Frequency
10.5-15-5	7	7
15.5-20.5	10	17
20.5-25.5	13	30
25.5-30.5	26	56 (c.f)
$(l_1)30.5 - 35.5$	35(f)	91
35.5-40.5	22	113
40.5-45.5	11	124
45.5-50.5	5	129
	N = 129	

Median, M = Size of $\left(\frac{N}{2}\right)$ th item; N = ∑f = 129

= Size of $\left(\frac{129}{2}\right)$ item

= Size of 64.5th item

64.5th item lies in 91st cumulative frequency of the series. The corresponding class interval, 30.5-35.5, would therefore, be the Median class interval.

Using the formula,

$$\mathrm{M} = l_1 + \frac{\frac{\mathrm{N}}{2} - \mathrm{c.f.}}{\mathrm{f}} \times i$$

$$= 30.5 + \frac{\frac{129}{2} - 56}{35} \times 5$$

$$= 30.5 + \frac{64.5 - 56}{35} \times 5 = 30.5 + \frac{8.5}{35} \times 5$$

= 30.5 + 1.2 = 31.7

Median = 31.7 marks.

(6) Median of the Series with Unequal Class Intervals

Illustration.

Calculate median of the following distribution of data:

Class Interval	0-5	5-10	10-20	20-30	30-50	50-70	70-100
Number of Students	12	15	25	40	42	14	8

Solution:

In such distributions, no specific treatment of the data is needed. Median can be calculated straight away as in the case of other frequency distributions.

Class Interval	Frequency (f)	Cumulative Frequency
0-5	12	12
5-10	15	27
10-20	25	52 (c.f.)
(l_1) 20-30	40 (f)	92
30-50	42	134
50-70	14	148
70-100	8	156
	N = 156	

M = Size of $(\frac{N}{2})$th item; N = 156

= Size of $(\frac{156}{2})$ th item

= Size of 78th item

This lies in 92th cumulative frequency and the corresponding Median class is 20-30.

$\therefore l_1 = 20$, c.f. $= 52$,f $= 40$ and $i = 10$

Substituting the values in the formula, we have

$$\text{M} = l_1 + \frac{\frac{\text{N}}{2} - \text{c.f.}}{\text{f}} \times i$$

$$= 20 + \frac{\frac{156}{2} - 52}{40} \times 10$$

$$= 20 + \frac{78 - 52}{40} \times 10 = 20 + \frac{26}{40} \times 10$$

= 20 + 6.5 = 26.5

Median = 26.5.

To Locate Missing Frequency Illustration.

Find the missing frequency in the following distribution if N= 100 and **M** = 30.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
Number of Students	10	?	25	30	?	10

Solution:

Two frequencies are missing and let the missing frequencies be denoted by f, and f_2. We need two equations, we will get one from summation of frequencies and one from median formula.

Marks	Frequency (f)	Cumulative Frequency
0-10	10	10
10-20	f_1	10 + f_1
20-30	25	35 + f_1
30-40	30	65 +f_1
40-50	f_2	65 + f_1 +'f_2

50-60	10	$75 + f_1 + f_2$
	N = 100	

1st Equation: From summation of frequencies,

$75 + f_1 + f_2 = 100 \Rightarrow f_1 + f_2 = 25$ ...(i)

2nd Equation: From information regarding median, M=30, Median class is 30-40.

Now,

$$M = l_1 + \frac{\frac{N}{2} - \text{c.f.}}{f} \times i$$

$$30 = 30 + \frac{50 - (35 + f_1)}{30} \times 10$$

$$0 = \frac{15 - f_1}{3} \quad \text{or} \quad f_1 = 15$$

Substituting $f_1 = 15$ from equation (ii), in equation (i) As $f_1 + f_2 = 25$

Put $f_1 = 15$

∴ $15 + f_2 = 25$

∴ $f_2 = 25- 15 = 10$

Thus, $f_1 = 15$, $f_2 = 10$.

Illustration.

Find the missing frequency in the following distribution if N is 60 and median is 40.

Marks	0-10	10-30	30-60	60-80	80-90
Frequency	5	f_1	f_2	8	2

Solution:

Marks	Frequency (f)	Cumulative Frequency
0-10	5	5
10-30	f_1	$5 + f_1$
30-60	f_2	$5 + f_1 + f_2$
60-80	8	$13 + f_1 + f_2$

80-90	2	$15 + f_1 + f_2$
	N = 60	

∴ N = 60

$\therefore 15 + \mathrm{f}_1 + \mathrm{f}_2 = 60$ or $\mathrm{f}_1 + \mathrm{f}_2 = 45$

Median Size = $\frac{N}{2} = \frac{60}{2}$ = 30th term

(∴ Median is 40, so Median class is 30-60)

$$\mathrm{M} = l_1 + \frac{\frac{\mathrm{N}}{2} - \mathrm{c.f.}}{\mathrm{f}} \times i$$

$$40 = 30 + \frac{30 - (5 + f_1)}{f_2} \times 30$$

=> $10 = \frac{(25 - \mathrm{f}_1)}{\mathrm{f}_2} \times 30$

=>$10\mathrm{f}_2 = 750 - 30\mathrm{f}_1$

$30\mathrm{f}_1 + 10\mathrm{f}_2 = 750$

Thus, we get

$f_1 + f_2 = 45$ ...(i)

$30\ f_1 + 10f_2 = 750$ ...(ii)

Dividing equation (ii) by 10, we get

$3\ f_1 + f_2 = 75$ ...(iii)

Now subtracting (i) from (iii), we get

$$3\mathrm{f}_1 + \mathrm{f}_2 = 75$$

$$\underset{-}{f_1} \pm \mathrm{f}_2 = \underset{-}{45}$$

$$2\mathrm{f}_1 = 30$$

=> $f_1 = 15$

Now substituting $f_1 = 15$ in (i), we get

$f_2 = 45 - 15 = 30$

Hence, $f_1 = 15$, $f_2 = 30$.

Graphic Determination of Median

Median value of a series may also be determined through the graphic presentation of the data in the form of ogives. This may be done in two ways.

(1) Presenting the data graphically in the form of less than’ or ‘more than’ ogives.

(2) Presenting the data graphically and simultaneously in the form of less than’ and ‘more than’ ogives. That is, the two ogives are superimposed upon each other to determine the median value.

(1) ‘Less than' or 'More than' Ogive Approach

According to this approach, a frequency distribution series is first converted into a ‘less than’ or ‘more than’ cumulative series as in the case of ogives. Data are presented graphically to make a ‘less than’ or ‘more than ogive, $(\frac{N}{2})$ th item of the series is determined and from this point (on the Y-axis of the graph) a perpendicular is drawn to the right to cut the cumulative frequency curve. The median value of the series is the one where the cumulative frequency curve is cut by the perpendicular corresponding to the X-axis. The following illustration explains the method.

Illustration.

Determine median value of the following series using graphic method:

Marks	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35^40
Number of Students	4	6	10	10	25	22	18	5

Solution:

Using ‘Less than’ ogive approach, the calculation of the median value involves the following steps:

(i) Convert the series into a 'less than’ cumulative frequency distribution and graph the data as under:

Estimation of the Median: ‘Less than’ Ogive Approach

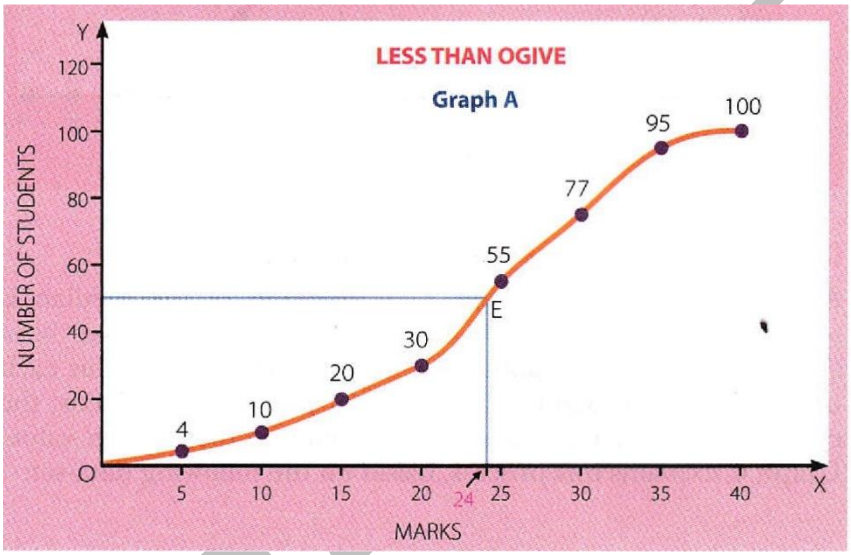
Marks	Cumulative Frequency
Less than 5	4
Less than 10	10
Less than 15	20
Less than 20	30
Less than 25	55
Less than 30	77
Less than 35	95
Less than 40	100

(Ans. Median = 24)

(ii) Find out $\left(\frac{N}{2}\right)$th item and mark it on the Y-axis. In the above illustration $\left(\frac{N}{2}\right)$th item is $\frac{100}{2} = 50$.

(iii) Draw a perpendicular from 50 to the right to cut the cumulative frequency curve (ogive) at point E.

(iv) From the point E where cumulative frequency curve (ogive) is cut, draw a perpendicular on X-axis. The point at which it touches X-axis will be the median value of the series, as shown in Graph A:



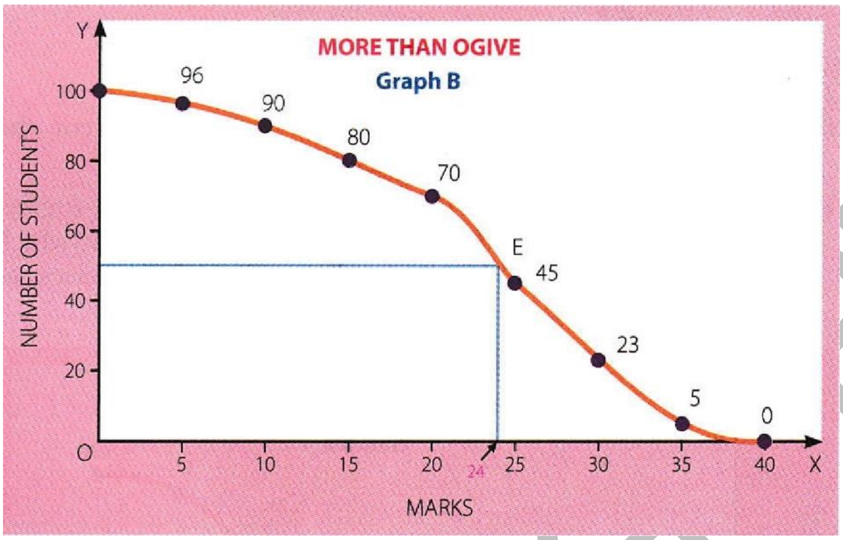
Using ‘more than’ ogive approach, the calculation of the median involves the similar procedure, as should be evident from the following table and Graph B:

Estimation of the Median: ‘More than’ Ogive Approach

Marks	Cumulative Frequency
More than 0	100
More than 5	96
More than 10	90
More than 15	80
More than 20	70
More than 25	45
More than 30	23
More than 35	5

More than 40	0

(Ans. Median = 24)



(2) 'Less than' and 'More than' Ogive Approach

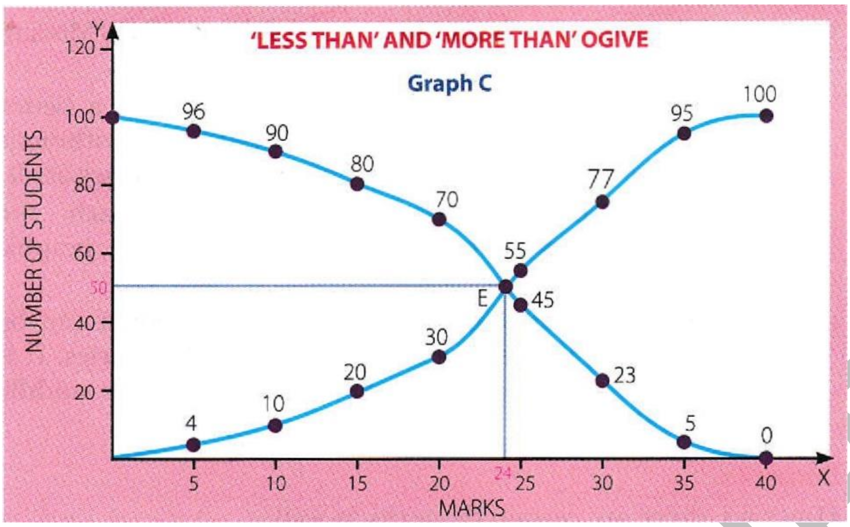
Another way of the graphic determination of the median is simultaneous graphic presentation $^{\%}$of both the ‘less than’ and ‘more than’ ogives. Mark the point E where the ogive curves cut each other; draw a perpendicular from that point on the X-axis; the corresponding value on the X-axis would be the median value. Graph C explains this process, based on the following data-set:

Estimation of the Median:

‘Less than’ and ‘More than’ Ogive Approach

Marks	Cumulative Frequency	Marks	Cumulative Frequency
Less than 5	4	More than 0	100
Less than 10	10	More than 5	96
Less than 15	20	More than 10	90
Less than 20	30	More than 15	80
Less than 25	55	More than 20	70
Less than 30	77	More than 25	45
Less than 35	95	More than 30	23
Less than 40	100	More than 35	5
		More than 40	0

(Ans. Median = 24)



Two Principal Merits of Median as the Measure of Central Tendency of the Series are

(i) Unlike arithmetic mean, median is free from the effect of extreme values of the series.

(ii) Median value can be estimated even in case of an incomplete statistical series.

What is the Principal Demerit of Median as the measure of Central Tendency?

That unlike arithmetic mean, the median value is not capable of further algebraic treatment, and that its estimation is not based on ail items of the series.

Merits and Demerits of Median as a Measure of Central Tendency

Merits

Median is a widely used measure of the central tendency, particularly in the field of socio-economic studies. Some of its main merits are as under:

(1) Simple: It is a very simple measure of the central tendency of the series. In the case of simple statistical series, just a glance at the data is enough to locate the median value.

(2) Free from the Effect of Extreme Values: Unlike arithmetic mean, median value is not distorted by the extreme values of the series.

(3) Certainty: Certainty is another merit of the median. Median value is always a certain specific value in the series.

(4) Real Value: Median value is a real value and is a better representative value of the series compared to arithmetic average, the value of which may not exist in the series at all.

(5) Graphic Presentation: Besides algebraic approach, the median value can be estimated also through the graphic presentation of data.

(6) Possible Even When Data is Incomplete: Median can be estimated even in the case of certain incomplete series. It is enough if one knows the number of items and the middle item(s) of the series.

Demerits

However, there are certain demerits as well:

(1) Lack of Representative Character: Median fails to be a representative measure in case of such series the different values of which are wide apart from each other. Also, median is of limited representative character as it is not based on all the items in the series.

(2) Unrealistic: When the median is located somewhere between the two middle values, it remains only an approximate measure, not a precise value.

(3) Lack of Algebraic Treatment: Arithmetic mean is capable of further algebraic treatment, but median is not. For example, multiplying the median with the number of items in the series will not give us the sum total of the values of the series.

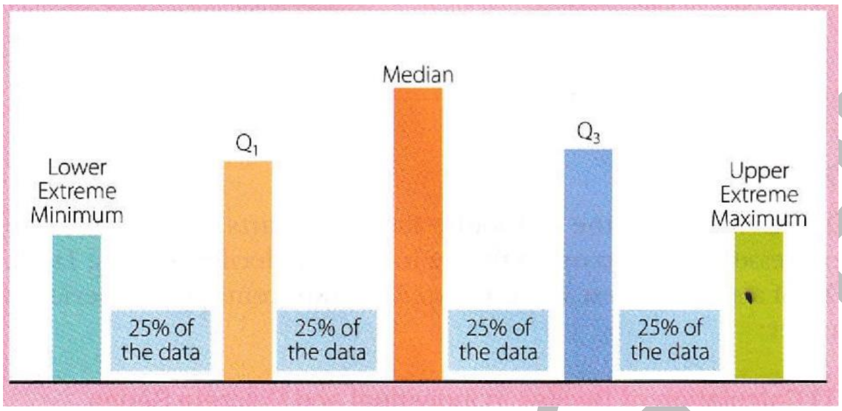
However, median is quite a simple method of finding an average of a series. It is quite a commonly used measure in the case of such series which are related to qualitative observations as IQ, and health of the students.

Methods of Calculating Median—A Glance

Type of Series	**Method of Calculation**
1. Individual Series	(a) M = Size of $\left(\frac{N+1}{2}\right)$th item (Here, N = Total of items.) (b) If $\left(\frac{N+1}{2}\right)$ comes to be in fractions, the median would be average of the two middle values of the series.
2. Discrete Series/ Frequency Array	M = Size of $\left(\frac{N+1}{2}\right)$ item (Here, N = Sum of the frequencies ∑f.)
3. Frequency Distribution Series	M = Size of $\left(\frac{N}{2}\right)$th item Median class corresponds to that cumulative frequency which includes the above value. Following formula is applied to determine actual median value: $\text{M} = l_1 + \frac{\frac{\text{N}}{2} - \text{c.f.}}{\text{f}} \times i$

2. PARTITION VALUE: QUARTILE

The value that divides the series into more than two parts is called Partition Value. If a statistical series is divided into four equal parts, the end value of each part is called a quartile. It is written as Q. The First Quartile or Q, is also known as lower quartile. The Second Quartile, or Q_2 is the same as Median of the series. The Third Quartile, or Q_3 is also called upper quartile.



Did you know?

The first quartile (Q_1 or lower quartile has 25% of the items of the distribution below it and 75% of the items above it. The second quartile (Q_2) or median has 50% of items below it and 50% of the items above it.

The third quartile (Q_3) or upper quartile has 75% of the items below it and 25% of items above it.

Estimation of Q_1 and Q_3

Quartile values (Q_1 and Q_3) are estimated differently for different sets of series, as under:

(1) Individual and Discrete Series

For the individual and discrete series, Q_1 and Q_3 are determined using the following formulae:

FORMULA

$Q_1 =$ Size of $\left(\frac{N+1}{4}\right)$th item of the series

$Q_3 =$ Size of $3\left(\frac{N+1}{4}\right)$th item of the series

In individual series, N = Number of items,

while in discrete series, N = Sum of frequencies, ∑f.

(2) Frequency Distribution Series

In frequency distribution series, the class interval of Q, and Q_3 are first identified as under:

Class interval of $Q_1 =$ Size of $(\frac{N}{4})$ th item

Class interval of $Q_3 =$ Size of $3(\frac{N}{4})$ th item

FORMULA

$$Q_1 = l_1 + \frac{\left[\frac{N}{4} - c.f.\right]}{f} \times i$$

$$Q_3 = l_1 + \frac{\left[3\left(\frac{N}{4}\right) - \text{c.f.}\right]}{\text{f}} \times i$$

Deciles

Deciles distribute the series into ten equal parts, and is generally expressed as D. Accordingly, we have nine deciles, D_1 D_2, D_3,..., **Dg,** of a series. These are estimated for different types of series as under:

(1) Calculation of Deciles in Individual and Discrete Series

Following formulae are used in the estimation of D_1, D_4 or D_9 respectively:

FORMULA

$$D_1 = \text{Size of} \left(\frac{N+1}{10}\right) \text{th item}$$

$$D_4 = \text{Size of } 4\left(\frac{N+1}{10}\right) \text{th item}$$

$$D_9 = \text{Size of } 9\left(\frac{N+1}{10}\right) \text{th item}$$

(2) Calculation of Deciles in Frequency Distribution Series

In frequency distribution series, first the class interval of the concerned decile is identified using the following formulae:

Class interval of D_1 = Size of $\left(\frac{N}{10}\right)$ th item

Class interval of D_4 = Size of $4\left(\frac{N}{10}\right)$ th item

Class interval of D_9 = Size of $9\left(\frac{N}{10}\right)$ th item

Actual value of the concerned decile is calculated using the following formulae:

FORMULA

$$D_1 = l_1 + \left[\frac{\frac{N}{10} - \text{c.f.}}{\text{f}}\right] \times \text{i}$$

$$D_4 = l_1 + \left[\frac{4\left(\frac{N}{10}\right) - c.f.}{f}\right] \times i$$

$$D_9 = l_1 + \left[\frac{9\left(\frac{N}{10}\right) - c.f.}{f}\right] \times i$$

Percentiles

Percentiles divide the series into 100 equal parts, and is generally expressed as P. There are 99 percentiles in a series ranging from

P_1, P_2, P_3, ... , T_{99}. Obviously P_{50} would be the median of the series as it is the centrally located value in the series. Likewise b˃25 is the same as Qi and P_{75} is the same as of the series. If a student has secured 89 percentile in an MBA entrance examination, it means that his position is below 11 per cent of total candidates appeared in the examination. Percentiles are estimated for different types of series as under:

(1) Estimation of Percentiles in Individual and Discrete Series

In the individual and discrete series, P_1, P_2, P_3, ... are estimated using the following formulae:

FORMULA

$$P_1 = \text{Size of} \left(\frac{N+1}{100}\right) \text{ th item}$$

$$P_4 = \text{Size of } 4\left(\frac{N+1}{100}\right) \text{ th item}$$

$$P_{gg} = \text{Size of } 99\left(\frac{N+1}{100}\right) \text{ th item}$$

(2) Estimation of Percentile in Frequency Distribution Series

In frequency distribution series, the class interval of P_1, P_2, P_3, are first identified using the following formulae:

$$\text{Class interval of } P_1 = \text{Size of} \left(\frac{N}{100}\right) \text{ th item}$$

$$\text{Class interval of } P_4 = \text{Size of } 4\left(\frac{N}{100}\right) \text{ th item}$$

$$\text{Class interval of } P_{99} = \text{Size of } 99\left(\frac{N}{100}\right) \text{ th item}$$

(That is, class interval of P_{99} corresponds to size of $99\left(\frac{N}{100}\right)$th item in the cumulative frequency of the series.)

Actual value of the concerned percentile is calculated using the following formulae:

FORMULA

$$P_1 = l_1 + \left[\frac{\frac{N}{100} - \text{c.f.}}{\text{f}}\right] \times i$$

$$P_4 = l_1 + \left[\frac{4\left(\frac{N}{100}\right) - \text{c.f}}{\text{f}}\right] \times i$$

$$P_{99} = l_1 + \left[\frac{99\left(\frac{N}{100}\right) - c.f.}{f}\right] \times i$$

The following table presents different partition values.

Table of Partition Values

Partition Values	Divisions of a Series	Number of Partition Values in Series
Quartile	4	3
Decile	10	9
Percentile	100	99

Illustration.

From the following data, calculate Q_1, Q_3, D_5 and P_{25}.

S. No. 21	15	40	30	26	45	50	54	60	65	70

Solution:

The data is first arranged in ascending order:

S. No.	X
1	15
2	21
3	26
4	30
5	40

6	45
7	50
8	54
9	60
10	65
11	70
N = 11	

$$Q_1 = \text{ Size of } (\frac{N+1}{4}) \text{ th item}$$

= Size of $(\frac{11+1}{4})$th item

= Size of 3rd item = 26 Q_3

= Size of 3rd item

= Size of $3(\frac{N+1}{4})$th item

= Size of $\left(\frac{11+1}{4}\right)$item = 60

D_5 = Size of 5 $(\frac{N+1}{10})$th item

= Size of $55(\frac{11+1}{10})$th item

= Size of 6th item = 45

P_{25} = Size of $25(\frac{\mathrm{N}+1}{100})$ item

= Size of 25 $(\frac{11+1}{100})$th item

= Size of 3rd item = 26

Thus, Q_1 = 26, Q_3 = 60, D_5 = 45, P_{25} = 26.

Illustration.

Calculate Q_1 Q_3, D_6 and P_{85} from the following data:

Size	10	11	12	13	14	15	16	17	18
Frequency	3	4	5	12	10	7	5	2	1

Solution:

Size	Frequency (f)	Cumulative Frequency
10	3	3
11	4	7
12	5	12
13	12	24
14	10	34
15	7	41
16	5	46
17	2	48
18	1	49
	N = 49	

$$Q_1 = \text{Size of } \left(\frac{N+1}{4}\right) \text{ th item } = \text{Size of } \left(\frac{49+1}{4}\right) \text{ th item}$$

Size of 12.5th item =13

$$Q_3 = \text{Size of } 3\left(\frac{N+1}{4}\right) \text{ th item } = \text{Size of } 3\left(\frac{49+1}{4}\right) \text{ th item}$$

Size of 37,5th item = 15

$$D_6 = \text{Size of } 6\left(\frac{N+1}{10}\right) \text{ th item } = \text{Size of } 6\left(\frac{49+1}{10}\right) \text{ th item}$$

= Size of 30th item = 14

$$P_{85} = \text{Size of } 85\left(\frac{N+1}{100}\right) \text{th item } = \text{size of } 85\left(\frac{49+1}{100}\right) \text{ th item}$$

= Size of 42.5th item

= 16 Thus, Q_1 = 13, Q_3 = 15, D_6 = 14, P_{85} = 16.

Illustration.

Calculate the values of Q_1, Q_3, **D_8** and **P_{50}** from the following data:

Wages (Rs.)	0-10	10-20	20-30	30-40	40-50
Number of Workers	22	38	46	35	19

Solution:

Calculation of $Q_{1?}$ Q_3, D_8 and P_{56}

Wages (Rs.)	Frequency (f)	Cumulative Frequency
0-10	22	22
10-20	38	60
20-30	46	106
30-40	35	141
40-50	19	160
	N = 160	

$$Q_1 = \text{Size of } (\frac{N}{4})\text{th item} = \text{Size of } (\frac{160}{4})\text{th item}$$

= Size of 40th item.

Q_1 lies in 60th cumulative frequency. The class interval corresponding to this cumulative frequency is 10-20.

$$Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f} \times i$$

$$= 10 + \frac{\frac{160}{4} - 22}{38} \times 10$$

$$= 10 + \frac{40 - 22}{38} \times 10 = 10 + \frac{18}{38} \times 10$$

= 10 + 4.74 = 14.74

= 10 + 4.74 = 14.74

$$Q_3 = \text{Size of } 3(\frac{N}{4}) \text{ th item} = \text{Size of } 3(\frac{160}{4}) \text{ th item}$$

= Size of 120th item

Q_3 lies in 141 th cumulative frequency. The class interval corresponding to this cumulative frequency is 30-40.

Thus, $Q_3 = l_1 + \frac{3(\frac{N}{4}) - \text{c.f.}}{\text{f}} \times i$

$$= 30 + \frac{3\left(\frac{160}{4}\right) - 106}{35} \times 10$$

$$= 30 + \frac{120-106}{35} \times 10 = 30 + \frac{14}{35} \times 10$$

= 30 + 4 = 34

$$\text{D}_8 = \text{SIZE OF } 8\left(\frac{\text{N}}{10}\right) \text{ TH ITEM}$$

$$= \text{ Size of } 8(\frac{160}{10}) \text{ th item}$$

= Size of 128th item

0_8 lies in 141th cumulative frequency. The class interval corresponding to this cumulative frequency is 30—40.

Thus, $\text{D}_8 = l_1 + \frac{8\left(\frac{\text{N}}{10}\right)-\text{c.f.}}{\text{f}} \times i$

$$= 30 + \frac{8\left(\frac{160}{10}\right)-106}{35} \times 10 = 30 + \frac{128-106}{35} \times 10$$

$$= 30 + \frac{22}{35} \times 10$$

= 30 + 6.29

= 36.29

$$P_{56} = \text{ Size of } 56(\frac{N}{100}) \text{ th item}$$

$$= \text{ Size of } 56(\frac{160}{100}) \text{ th item}$$

= Size of 89.6th item

P_{56} lies in 106th cumulative frequency. The class interval corresponding to this cumulative frequency is 20-30.

Thus, $\text{P}_{56} = l_1 + \frac{56(\frac{\text{N}}{100})-\text{c.f.}}{\text{f}} \times i$

$$= 20 + \frac{56(\frac{160}{100})-60}{46} \times 10$$

$$= 20 + \frac{89.6-60}{46} \times 10$$

$$= 20 + \frac{29.6}{46} \times 10$$

= 20 + 6.43

= 26.43

Thus, $Q_1 = 14.74$, $Q_3 = 34$, $D_8 = 36.29$, $P_{56} = 26.43$.

3. MODE

Mode is another important measure of central tendency of statistical series. It is the value which occurs most frequently in the series; that is, modal value has the highest frequency in the series. For example, if out of 100 students in a class, 70 students record their age as 15 years, the modal age of the students would be 15 years. The word 'Mode' is derived from the French word la Mode which means a thing in vogue or fashion. In statistics, it is generally represented by the letter Z.

Definition

According to **Kenny,** “The value of the variable which occurs most frequently in a distribution is called the mode. ”

Croxton and **Cowden** state, “The mode may be regarded as the most typical of a series of value. ”

Calculation of Mode

Mode is calculated differently for different types of series.

(1) Calculation of Mode in Individual Series

There are two ways of calculating mode in individual series:

(i) By Inspection.

(ii) By Converting Individual Series into Discrete Frequency Series.

(i) Mode by Inspection: This method involves just an inspection of the series. One is to simply identify the value that occurs most frequently in the series. Such a value is called mode.

Illustration.

Age of 15 students of a class is reported below. Find the modal age.

Age (Years)	22	24	17	18	17	19	18	21 20	21	20	23	22 22	22

Solution:

Arrange the series in an ascending order as:

Age (Years)	17	17	18	18	19	20	20	21	21	22	22	22	22	23	24

An inspection of the series shows that 22 occurs most frequently in the series.

Hence, Mode (Z) = 22.

(ii) By Converting Individual Series into Discrete Frequency Series: When the number of items in a series is very large, Inspection Method of finding out the mode becomes very difficult, if not impossible. In such cases, individual series are first converted into discrete frequency series or frequency array. The mode is then identified as the value corresponding to which there is highest frequency.

Illustration.

The table below presents death rate of population across different countries. Find the mode.

Death Rate (per thousand)	11.1	10.9	10.7	11.1	10.6	11.3	10.6
	10.7	10.6	10.9	10.6	10.5	10.4	10.6

When to use Inspection Method to identify Modal Value of the Series?

Inspection method is to be used only in case of such series where different items show different frequencies.

Solution:

First, we convert the series into a discrete frequency distribution in ascending order as under.

Death Rate	10.4	10.5	10^6	10.7	10.9	11.1	11.3
Frequency	1	1	5	2	2	2	I

10.6 is the value with highest frequency of 5. Hence,

Mode (Z) = 10.6.

(2) Calculation of Mode in Discrete Series or Frequency Array

There are two methods for calculation of mode in discrete frequency series: (i) Inspection Method; and (ii) Grouping Method.

(i) Inspection Method

As in the case of individual series, Inspection Method is sometimes used for the calculation of mode in discrete frequency series as well. An inspection of the series would show the mode or the value that has the highest frequency in the series, provided the series are regular and homogeneous.

Illustration.

Find out mode of the following series:

Income (Rs.)	110	120	130	140	150	160
Number of Persons	2	4	8	10	5	4

Solution:

A glance at the series reveals that 140 commands the highest frequency of 10 in the series.

Hence, 140 is the Mode (Z) of the series.

(ii) Grouping Method

In the case of discrete frequency distribution, inspection method is possible only when there is regularity and homogeneity in the series. However, the series may sometimes not be regular or homogeneous.

More than one value may command the highest frequency in the series. In such cases, Grouping Method of the calculation of mode is used. The method involves the construction of: (a) Grouping Table, and (b) Analysis Table. The illustration below explains the calculation of mode using the grouping technique.

Illustration.

Given the following data, calculate mode using the grouping technique.

Size	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	3	8	10	12	16	14	10	8	17	5	4	1

Solution:

Table 1. Grouping Table for the Estimation of Mode

Size	**(I)** **Frequency**	**(H)** **(1+2)**	**(III)** **(2+3)**	**(IV)** **(1 + 2+3)**	**(V)** **(2+3+4)**	**(VI)** **(3+4+5)**
2	3		8 + 10 = 18 12 + 16 =28	3 + 8 + 10 = 21	8 + 10 + 12 = 30	10 + 12 + 16 =38
3	8	3 + 8 – 11	14 + 10 = 24	12 + 16 + 14 =42	16+ 14 + 10 =40	14 + 10 + 8 = 32
4	10	10 + 12 = 22	8+ 17 =25	10 + 8 + 17= 35		
5	12		5 + 4 = 9		8 + 17 + 5 = 30	

6	J 6	16 + 14 =30 10 + 8 = 18 17 + 5 = 22 4+ 1 = 5		5 + 4+1 = 10		17 + 5 + 4 = 26
7	14					
8 9	10 8					
10	17					
11	5					
12	4					
13	1					

Check the following Tables 1 and 2 and the subsequent details.

There are 6 columns in the table on the frequency and its groupings, as in Table 1.

In Column I, the frequencies as given in the question are noted against the respective items.

In Column II, the frequencies are grouped in twos beginning with the 1st item.

In Column III, the frequencies are grouped in twos beginning with 2nd item.

In Column IV, the frequencies are grouped into threes beginning with 1st item.

In Column V, the frequencies are grouped into threes beginning with 2nd item.

In Column VI, the frequencies are grouped into threes beginning with 3rd item.

Highest value in all the columns are shown in bold letters or are underlined, or encircled.

Table 2. Analysis Table

Column	**Size of items containing maximum frequency**											
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
I										✓		
II					✓	✓						

III				✓	✓							
IV				V	✓							
V					✓	✓	✓					
VI			✓	✓	✓							
Total	—	—	1	3	5	3	1	—	1	—	—	—

Analysis Table shows those items corresponding to which there are highest frequencies in different columns. Thus, in column I the highest total is 17 and this corresponds to the item 10. For column I a tick mark (V) is accordingly noted against 10 in the Analysis Table. Likewise, in column II the highest total is 30 and this corresponds to the items 6 and 7 (bracketed). A tick mark (’Q is accordingly noted against 6 as well as 7 in the Analysis Table. Similarly, in column III the highest total is 28 and it corresponds to the items 5 and 6. Thus, a tick mark is noted against 5 and 6. In column IV, the highest total is 42 and it corresponds to the items 5, 6 and 7. Thus, a tick mark is noted against 5, 6, and 7 in column IV. Likewise, tick marks are noted in other columns. The Analysis Table shows that item 6 has the largest number of tick marks. Thus, the Mode (Z) = 6.

(3) Calculation of Mode in Frequency Distribution Series

As in the case of discrete frequency series, calculation of mode in continuous series has two methods: (i) Inspection Method, and (ii) Grouping Method.

However, in the case of continuous series, we first find out the modal class, using either of these two methods, and afterwards, calculate the exact value of mode with the aid of the following formula:

FORMULA

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

(Here, Z = Value of the mode; l_1 = lower limit of the' modal class; I', = The frequency of the modal class; f_1= The frequency of pre-modal class; f_0 = Frequency of the next higher class or post-modal class; i = Size of the modal group.)

Below we discuss with illustrations the application of two methods for the calculation of mode in continuous series.

You Must Understand it

That in case of frequency distribution series, inspection method only helps in identification of the modal class interval, not the exact modal value.

(i) Inspection Method

When frequencies of a continuous series increase and decrease in some systematic order, the modal class may be found merely by an inspection of the series. The following illustration is an example of such series where modal class is found by inspection. Having found the modal class, the actual value of the mode is calculated using the formula given above.

Illustration.

Calculate mode from the following data:

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	2	5	7	5	2

Solution:

Calculation of Mode

Class Interval	Frequency
0-10	2
10-20	5 (f_0)
(l_1) 20-30	7 (f_1)
30-40	5 (f_2)
40-50	2

A glance at the series reveals that 20-30 is the modal class because it has the maximum frequency, i.e., 7. Following formula is used to calculate modal value:

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 20 + \frac{7-5}{2(7) - 5 - 5} \times 10$$

$$= 20 + \frac{2}{14 - 10} \times 10$$

$$= 20 + \frac{2}{4} \times 10$$

=20 + 5-25

Mode (Z) = 25.

(ii) Grouping Method

Inspection method of finding out modal class fails in case the frequencies of the series are not increasing or decreasing in some systematic pattern. It fails also, when there are more than one class intervals with higher frequencies in the distribution. In such cases, modal class is determined using the grouping method.

We have already discussed the grouping method in the case of discrete frequency series; it is the same in the case of continuous series as well. After the modal class has been identified, the actual value of mode is calculated using the same formula, viz.

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Solution:

Grouping of the frequencies is essential in this series, because there are two class intervals, i.e., 10-15 and 20-25 with highest frequency, 10.

Hence, the grouping table and the subsequent Analysis Table are as follows:

Illustration.

Calculate mode of the following series:

Class Interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Frequency	1	2	10	4	10	9	2

Grouping Table

Class Interval (I)		(II) (1+2)	(III) (2+3)	(IV) (1+2 + 3)	(V) (2+3+4)	(VI) (3+4+5)
0-5	1	1+2 = 3		1 + 2 + 10 = 13		
5-10	2		2 + 10 = 12		2 + 10 + 4 = 16	
10-15	10	10 + 4 = 14				10 + 4+10 =24
15-20	4		4+10 =14	4 + 10 + 9 =23		
20-25	10	10 + 9 =19			10 + 9 + 2 = 21	
25-30	9		9 + 2 = 11			
30-35	2					

Analysis Table

Column	Class intervals corresponding to highest frequencies in the Grouping Table						
	0-5	5-10	10-15	15-20 20-25		25-30	30-35
I			✓				
11						✓	
III				✓			
IV				✓	✓	✓	
V					✓	✓	✓
VI			✓	✓	✓		
Total	—	—	2	3	6	3	1

Analysis Table shows that of all the class intervals, 20-25 has the highest frequency (6). It has got the maximum (S) marks. Accordingly, 20-25 is the modal class interval of the series. The actual value of mode is,

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

l_1— lower limit of modal group 20

f_1 = frequency of modal class =10

f_0= frequency of class preceding the modal class (15-20) = 4

f_2= frequency of class following the modal class (25-30) = 9

i = class interval = 25 - 20 =5

Hence, $Z = 20 + \frac{10-4}{20-4-9} \times 5$

$$= 20 + \frac{6}{7} \times 5$$

= 20 + 4.29 = 24.29

Mode (Z) = 24.29

(4) Calculation of Mode when Class Intervals are Unequal

In case of unequal class intervals, mode will be determined by a different method. To calculate mode, unequal class intervals are made equal, as in the following illustration.

Illustration.

Class Interval	0-4	4-8	8-12	12-14	14-16	16-20	20-24	24-26	26-32	32-34	34-40	40-44	44-48
Frequency	2	4	7	5	8	3	2	1	13	5	3	3	9

Solution:

Change unequal class intervals into equal class-intervals and work out an analysis table, as under:

Class Interval	(I) Frequency	(ii) (1 + 2)	(III) (2+3)	(IV) (1+2+3)	(V) (2+3+4)	(VI) (3+4+5)
0-8	2+4=6 f_0	]6 + 20 = 26		]6 + 20 + 5 = 31		
8-16	7+5 + 8=20 f_1		]20 + 5 =25		]20 + 5+14 =39	
16-24	3+2=5 f_2	]5 + 14 = 19				]5+14 + 8 =27
24-32	1 + 13=14		]14 + 8 = 22	]14 + 8+12 =34		
32-40	5+3=8	]8 + 12 = 20				
40-48	3+0=12					

Analysis Table

Column	0-8	8-16	16-24	24-32	32-40	40-48
I		✓				
II	✓	✓				
III		✓	✓			

IV				✓	✓	✓
V		✓	✓	✓		
VI			✓	✓	✓	
Total	1	4	3	3	2	1

Mode is in 8-16 class interval. By using the formula,

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$l_1 = 8, f_1 = 20; f_0 = 6; f_2 = 5; i = 8$$

$$z = 8 + \frac{20 - 6}{2(20) - 6 - 5} \times 8$$

$$= 8 + \frac{14}{40 - 11} \times 8 = 8 + \frac{14}{29} \times 8$$

Z = 8 + 3.86

= 11.86

Mode (Z) = 11.86.

Illustration.

Calculate mode of the following data:

Marks	0-5	5-10	10-20	20-40	40-60	60-80	80-90	90-100
Number of Students	5	7	9	25	30	24	8	6

Solution:

This is a series with unequal class intervals. Calculation of mode requires that this series be converted into one with equal class interval of 20, as shown below:

Marks	**Number of Students**
0-20	5 + 7+9 = 21
20-40	25 = 25 (f_0)

(l_1) 40-60	30 = 30 (f_1)
60-80	24 = 24 (f_2)
80-100	8 + 6 = 14

A glance at the table reveals that 40-60 is the modal class. The actual value of mode is,

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 40 + \frac{30 - 25}{2(30) - 25 - 24} \times 20$$

$$= 40 + \frac{5}{60 - 49} \times 20$$

$$= 40 + \frac{5}{11} \times 20 = 40 + 9.09 = 49.09$$

Mode (Z) = 49.09.

(5) Calculation of Mode in Case of Inclusive Series Illustration.

Calculate mode of the following series:

Class Interval	10-19	20-29	30-39	40-49	50-59	60-69	70-79
Frequency	10	12	18	30	16	6	8

Solution:

For the calculation of mode, inclusive series are first to be converted into exclusive series, as under:

Class Interval	Exclusive Class Intervals	Frequency
10-19	9.5-19.5	10
20-29	19.5-29.5	12
30-39	29.5-39.5	18 (f_0)
40-49	(l_1) 39.5-49.5	30 (f_1)

50-59	49.5-59.5	16 (f_2)
60-69	59.5-69.5	6
70-79	69.5-79.5	8

Note

That for the calculation of mode in case of inclusive series, the series must be converted into exclusive as a first step.

A glance at the above table reveals that 39.5-49.5 is the modal class interval. The actual value of mode is,

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 39.5 + \frac{30 - 18}{2(30) - 18 - 16} \times 10$$

$$= 39.5 + \frac{12}{60 - 34} \times 10$$

$$= 39.5 + \frac{12}{26} \times 10$$

= 39.5 + 4.62 = 44.12

Mode (Z) = 44.12.

(6) Calculation of Mode when Only Mid-values of the Class Intervals are Known

Illustration.

Calculate mode, given the following data-set:

Mid-value	15	25	35	45	55	65	75	85
Frequency	5	8	12	16	28	15	3	2

Solution:

A series with ‘mid-value’ is first to be expanded as a series with class intervals, as under:

Expansion of Mid-values as Class Intervals for the Estimation of Mode

Mid-value	Class Interval	Frequency
15	10-20	5

25	20-30	8
35	30-40	12
45	40-50	16 (f_1)
55	(l_1) 50-60	28 (f_1)
65	60-70	15
75	70-80	3
85	80-90	2

Class interval 50-60 is the one with highest frequency. This is the modal class interval, as the frequencies are increasing or decreasing in a systematic pattern. The actual value of mode is,

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 50 + \frac{28 - 16}{2(28) - 16 - 15} \times 10$$

$$= 50 + \frac{12}{56 - 31} \times 10$$

$$= 50 + \frac{12}{25} \times 10$$

= 50 + 4.8 = 54.8

Calculation of Mode in 'Less than' Cumulative Frequency Distribution

Illustration.

Find out mode of the following series:

Wages (Rs.) less than	200	300	400	500	600	700	800	900
Number of Workers	5	18	38	70	90	95	98	100

Solution:

For the calculation of Mode, ‘Less-than’ cumulative frequency distribution is first to be converted into a normal continuous frequency distribution, as under:

Wages **(Rs.)**	**Number of Workers** **(f)**
100-200	5

200-300	18- 5 = 13
300-400	38- 18 = 20 (f_0)
(l_1) 400-500	70 - 38 = 32 (f_1)
500-600	90 - 70 = 20 (f_2)
600-700	95 - 90 = 5
700-800	98 - 95 = 3
800-900	100-98 = 2

An inspection of the series reveals that 400-500 is the modal class. The value of mode is,

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 400 + \frac{32 - 20}{2(32) - 20 - 20} \times 100$$

$$= 400 + \frac{12}{64 - 40} \times 100$$

$$= 400 + \frac{12}{24} \times 100$$

$$= 400 + 50 = 450$$

Mode (Z) = 450.

Note

That in case of normal distributions, mean, median and mode of the series tend to coincide.

Calculation of Mode through Arithmetic Mean and Median

One can calculate mode of a series through mean and median of that series. The formula is:

FORMULA

$$Z = 3M - 2\overline{X}$$

(Here, Z =Mode; M = Median; and X = Mean.)

Here, it may be noted that in the case of normal distributions, or symmetrical distributions, mean, median and mode coincide with each other. Thus, this formula is most useful in this case. However, in asymmetrical distribution the formula may also be used for the calculation of mode.

Illustration.

Calculate mode of a series, the mean and median values of which are 16 cm and 20 cm respectively.

Solution:

Mode = 3 Median - 2 Mean

Or, $Z = 3M - 2\overline{X}$

Here, M = 20 cm, $\overline{X} = 16$ cm

$Z = 3 \times 20 - 2 \times 16$

Thus, Mode (Z)= 28 cm.

= 60-32 = 28

Graphic Method of Locating Mode

Mode of a series may be located through a graphic presentation of the data. It involves the following steps:

(i) Present the given information in the form of a histogram. Identify the highest rectangle. This corresponds to modal class of the series.

(ii) Join the top corners of the modal rectangle with the immediate next corners of the adjacent rectangles. The joining lines must cut each other.

(iii) The point where the joining lines cut each other points to the modal value. This value is determined by drawing a perpendicular from that point on to the X-axis. Illustration below should explain this process.

Illustration.

Calculate mode of the following series, using the Graphic Technique:

Expenditure	Number of Families
0-10	14
10-20	23
20-30	27
30-40	21
40-50	15

Solution:

The adjoining graph shows expenditure on the X-axis and number of families on the Y-axis. As is evident, class interval 20-30 corresponds to the highest rectangle in the graph, and is therefore, the modal class interval.

For the calculation of modal value, AD and BC lines are drawn to join the top ends of the modal rectangle with the immediate next ends of the adjoining rectangles. These lines are crossing each other at point E. A perpendicular is drawn from point E on the X-axis to locate the exact value of mode which is 24. A counter check of this modal value may be made by calculating mode with the following formula:

$$Z = l_1 + \frac{f_1 - t_0}{2f_1 - f_0 - f_2} \times i$$

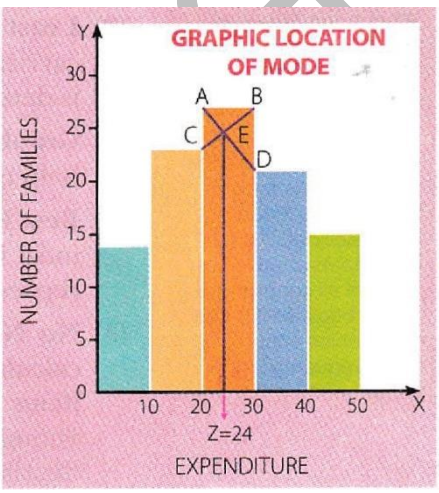
$$= 20 + \frac{27 - 23}{2(27) - 23 - 21} \times 10$$

$$= 20 + \frac{4}{54 - 44} \times 10$$

$$= 20 + \frac{4}{10} \times 10$$

$$= 20 + 4 = 24$$

Thus, Mode (Z) = 24.



Merits and Demerits of Mode as a Measure of Central Tendency

Merits

Mode, as a measure of central tendency has the following main merits:

(1) Simple and Popular: Mode is a very simple measure of the central tendency. Sometimes, just a glance at the series is enough to locate the modal value. Because of its simplicity, it is a very popular measure of central tendency.

(2) Least Effect of Marginal Values: Compared to mean, mode carries the least effect of marginal values in the series. Mode is determined only by the value with the highest frequencies.

(3) **Graphic Location**: Mode can be located graphically, with the help of histogram.

(4) Best Representative Value: Mode is that value which occurs most frequently in the series. Accordingly, mode is the best representative value of the series.

(5) No Need of knowing all the items or Frequencies: The calculation of mode does not require knowledge of all the items and frequencies of a distribution. In simple series, it is enough if one knows the item with highest frequency in the distribution.

Demerits

Some of the demerits of mode as a measure of central tendency are as under:

(1) Uncertain and Vague: Mode is an uncertain and vague measure of central tendency.

(2) Not Capable of Algebraic Treatment: Unlike mean, mode is not capable of further algebraic treatment.

(3) **Difficult**: When frequencies of all items are identical, it is difficult to identify the modal value.

(4) Complex Procedure of Grouping: Calculation of mode involves cumbersome procedure of grouping the data. If the extent of grouping changes, there will be a change in the modal value.

(5) Ignores Extreme Marginal Frequencies: It ignores extreme marginal frequencies. To that extent, modal value is not a representative value of all the items in a series.

Besides, one can question the representative character of the modal value as its calculation does not involve all items of the series.

Uses of Mode as a Measure of Central Tendency

Owing to simplicity of its calculation and understanding, mode is becoming an increasingly popular measure of the central tendency. In the trading sector, this measure is frequently used. Average expenditure, average income of the customers or popularity of goods in the market invariably refers to modal value. Likewise, in Meteorological Department, average rainfall or temperature of a place invariably refers to the modal value. In other spheres of life as well, the use and application of mode is gaining popularity.

What is the Principal Merit of mode as a Measure of Central Tendency?

It is the fact that sometimes just a glance at the series is enough to locate the modal value. Besides, it is the best representative value of all items of the series, because it is that value which occurs most frequently in the series,

The Principal Demerit of Mode as a Measure of Central Tendency

That it is difficult to identify modal value in case series are not regular or when different class intervals in a series show the same frequency.

Calculation of Mode—A Glance

Types of Series	Method of Calculation
1. Individual Series	(a) The value that occurs the most in the series is identified as mode of the series by inspection method. (b) If the frequency of all values is equal the same are changed into discrete frequency distribution and then mode is calculated.
2. Discrete Series/ Frequency Array	(a) The value that occurs the most in the series is identified by inspection method. (b) If items with highest frequency- are more than one, then grouping method is used.
3. Frequency Distribution Series	(a) Exclusive: Series with highest frequency is called modal class. The actual value of mode is determined using the following formula: $Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$ (b) Inclusive: Inclusive series is converted into exclusive series.
4. Moderately Asymmetrical Distributions	$Z = 3M - 2\overline{X}$

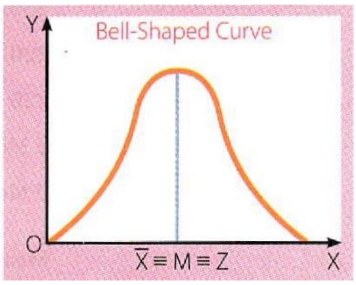
4. RELATIONSHIP AMONG MEAN, MODE AND MEDIAN

Before knowing the relationship among mean, median and mode, it is necessary to know whether the distribution is symmetrical or asymmetrical.

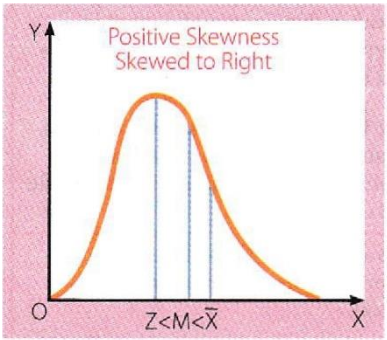
(i) In case of symmetrical distribution, mean, median and mode will be identical and frequency curve will be bell-shaped.

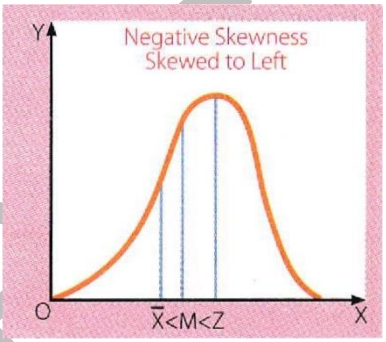
(ii) In case of asymmetrical distribution, value of mean, median and mode will be different. As such, frequency curve will not be bell-shaped. It may be skewed either^ to the right or to the left.

If the distribution is skewed more to the right, i.e., positive, then mean (X) and median (M) will be greater than mode (Z), i.e., (X, M > Z). In other words, mode (Z) is the minimum. If the distribution is skewed more to the left, i.e., negative, then mean



and median will be less than mode, < Z). In other words, mode (Z) will be maximum.





The following formula generalises the relationship between mode, median and mean of a series:

Mode = 3 Median - 2 Mean $Z = 3M-2\bar{X}$

Following example illustrates this relationship:

Illustration.

If in an asymmetrical distribution, median is 280 and mean is 310, what will be the mode?

Solution:

Mode = 3 Median - 2 Mean Mode = 3 × 280-2 × 310

= 840-620 = 220

When Mode is 220 and Median is 280, Mean would be:

220 = 280 × 3 - 2 Mean

220 = 840-2 Mean

2 Mean = 840 - 220

2 Mean = 620

Mean = $\frac{620}{2}$ = 310

When Mode is 220 and Mean is 310, Median would be:

220 = 3 Median-2x310

220 = 3 Median - 620

3 Median = 220 + 620

3 Median = 840

Median = $\frac{840}{3}$ = **280**

5. WHICH IS THE BEST AVERAGE?

There are many measures of central tendency like arithmetic mean, median, mode, etc. Which of these is the best average? There is no clear-cut answer to it. Different averages are suitable for different situations.

However, while selecting the relevant average, the following points must be kept in mind:

(1) **Objective:** Selection of average must conform to the objective of study. For instance, if all values are to be given equal importance, arithmetic mean will be most appropriate. However, if the value occurring most frequently in a series is to be identified, mode will be most relevant.

(2) **Number of Variables: In** case the number of variables in a series is very small, arithmetic mean is the best measure of the central tendency of the series.

(3) **Distribution of Items and Frequency:** If the value of large number of items in a series is small, but that of one or two items is large, then arithmetic mean may not be useful. If most of the values are located at the middle of the series or are related to qualitative facts, then the use of median should be the best option.

(4) **Importance to the Highest and the Lowest Items:** If no importance is to be attached to the highest and the lowest items of a series, then the use of median or mode should be most suitable.

(5) **Types of Series:** If in a series large number of items are similar to each other then the use of mode is not suitable.

6. COMPARATIVE FEATURES

Following comparative features must be kept in view while making use of different measures of central tendency:

(i) Mean and median can always be estimated with certainty but not the mode.

(ii) Median and mode are located in the middle of the frequency distribution, the mean may not be.

(iii) It is difficult to calculate mode as compared to mean and median.

(iv) Median and mode are not based on all the items of the distribution but mean is based on all the items of the distribution.

(v) Arithmetic examination of mean is possible. But it is not possible in case of median and mode.

(vi) Selection of a sample influences mean but not the median and mode.

(vii) Mean is influenced by the highest and the lowest items of the series. However, median and mode are not influenced by them.

(viii) Median and mode can be graphically located, but not the mean unless it is a situation of normal distribution.

Multiple Choice Questions

Select the correct alternative:

1. Median divides a series into how many parts?

(a) Two

(b) Three

(c) Four

(d) All of these

2. Which of the following is a kind of partition value?

(a) Arithmetic mean

(b) Median

(c) Quartile

(d) Both (b) and (c)

3. Which of the following formulae is used to find out median?

(a) (a) $M = l_1 + \frac{\frac{N}{4} - C.f.}{f} \times i$

(b) $M = l_2 + \frac{\frac{N}{4} - c.f.}{f} \times i$

(c) $M = l_1 + \frac{N - C.f.}{f} \times i$

(d) None of these

4. Dividing a series into ten equal parts is called:

(a) decile

(b) quartile

(c) percentile

(d) none of these

5. For calculating median, all items of the series are arranged in:

(a) descending order

(b) ascending order

(c) ascending or descending order

(d) none of these

6. Quartile is a type of:

(a) mathematical average

(b) statistical average

(c) partition value

(d) none of these

7. Mode refers to that value of a series that occurs ___________ times in the series.

(a) zero

(b) infinite

(c) maximum

(d) minimum

8. What is the relationship between mode, mean and median?

(a) $Z = 3M + 2\bar{X}$

(b) $z = 3M - 2\bar{X}$

(c) $\bar{x} = \frac{3M-Z}{2}$

(d) Both (b) and (c)

9. Out of the following, by which method mode can be calculated?

(a) Inspection method

(b) Grouping method

(c) Both (a) and (b)

(d) None of these

10. Which of the following formulae is used to find out P_{77} in frequency distribution?

(a) $l_1 - \frac{77\left(\frac{N}{10}\right)-\text{c.f.}}{f} \times i$

(b) $l_1 + \frac{77\left(\frac{N}{10}\right)-c.f.}{f} \times i$

(c) $l_1 + \frac{77\left(\frac{N}{10}\right)-c.f.}{f} \times i$

(d) $l_1 + \frac{77\left(\frac{N}{100}\right)-\text{c.f.}}{f} \times i$

11. Median of these numbers: 3, 5, 7, 9, 12 is:

(a) 3

(b) 6

(c) 7

(d) 12

12. Formula of D_7 in individual series is:

(a) Size of $\left(\frac{N+1}{100}\right)$th item

(b) Size of 7 $\left(\frac{N+1}{10}\right)^{\text{th}}$ item

(c) Size of 7 $\left(\frac{N}{10}\right)$th item

(d) none of these

13. In a distribution, the value around which the items tend to be most heavily concentrated is called:

(a) Median

(b) Mean

(c) Third quartile

(d) Mode

14. The sum of absolute deviations is minimum when taken from.

(a) Arithmetic Mean

(b) Median

(c) Mode

(d) None of these

15. The values of extreme items do not influence the average in case of

(a) Median

(b) Mean

(c) Mode

(d) None of these

16. Median:

(a) Can be determined graphically

(b) Not affected by extreme items

(c) Cannot be determined

(d) Involves complex calculations

17. Which of the following is not as measure of central tendency?

(a) Mean

(b) Median

(c) Mode

(d) Range

18. The second quartile is known as

(a) Median

(b) lower quartile

(c) Upper quartile

(d) None of these

19. The number of observations smaller than is the same as the number larger than it.

(a) Median

(b) Mode

(c) Mean

(d) None of these

20. Graphic location of mode is done with reference to:

(a) Cumulative frequency curve

(b) Frequency Polygon

(c) Frequency Curve

(d) Histogram

21. The most widely average used is:

(a) Arithmetic mean

(b) Median

(c) Mode

(d) Geometric mean

22. For open-end classification, which of the following is the best measure of central tendency

(a) Arithmetic Mean

(b) Geometric Mean

(c) Median

(d) Mode

23. What do you call the partition value which divides the series into two equal pats?

(a) Upper Quartile

(b) Median

(c) Mode

(d) Lower Quartile

24. The middle most value of a set of observations is

(a) Median

(b) Mode

(c) Mean

(d) None of these

25. Mode of a series is.

(a) An average value

(b) A middle value

(c) Highest frequency value

(d) None of above.

26. If mean of a series is 32 and median is 40, what would be the value of mode?

(a) 54

(b) 58

(c) 56

(d) 38

27. Most frequent occurring value in a series is called:

(a) Mode

(b) Median

(c) Mean

(d) Quartiles

28. Median can be calculated from series:

(a) Individual

(b) Discrete

(c) Continuous

(d) All the above

29. A distribution with more than two modes is called

(a) Bi-modal

(b) Multi-modal

(c) None of these

(d) Uni-modal

30. In a frequency distribution of a large number of values, the mode is:

(a) Smallest value

(b) Largest observation

(c) Observation with maximum frequency

(d) Maximum frequency of an observation

31. To find the median (mode), it is necessary to arrange the data in:

(a) Descending order

(b) Ascending order

(c) Ascending or descending order

(d) Any of the above

32. For the observations 5, 3, 6, 3, 5,10, 7, 2, there are modes.

(a) 2

(b) 3

(c) 4

(d) 5

33. A grouping table has

(a) 4 columns

(b) 6 columns

(b) 8 columns

(d) None of these

34. The value of a variate that occur most often is called

(a) Median

(b) Mean

(c) Mode

(d) None of these

35. In case of an even number of observations, which of the following is median?

(a) Any of the two middle-most value

(b) The simple average of these two middle values

(c) The weighted average of those two middle values

(d) Any of these

36. 50% of actual values will be below and 50% of will be above

(a) Mode

(b) Median

(c) Mean

(d) None of these

37. Mode of 0, 3, 5, 7, 9,12, 3 is:

(a) 6

(b) 0

(c) 3

(d) 5

38. A frequency distribution having two modes is said to be:

(a) Unimodal

(b) Bimodal

(c) Trimodal

(d) without mode

39. Histogram is useful to determine graphically the value of:

(a) Mean

(b) Median

(c) Mode

(d) All the above

40. Half of the number in an ordered set have values less than:

(a) Mean

(b) Median

(c) Mode

(d) None of these

41. If mode is ill defined then it is calculated with the help of formula:

(a) Mode = 2 Median – 3 mean

(b) Mode = 2 Median + 3 mean

(c) Mode = 3 Median + 3 mean

(d) Mode = 3 Median - 2 mean

42. c.f. is used for:

(a) Common factor

(b) Cumulative frequency

(c) Common value

(d) None of above

43. When the distribution is symmetrical, mean, median and mode

(a) Coincide

(b) Do not coincide

(c) Both

(d) None of these

44. Median of 2, 5, 8, 4, 9, 6.7, is:

(a) 9

(b) 8

(c) 8

(d) 6

45. In case of continuous frequency distribution, the size of __________ indicates class-interval in which the median lies.

(a) $\left[\frac{N+1}{2}\right]^{th}$ item

(b) $\left[\frac{N}{2}\right]^{th}$ item

(c) $\left[\frac{N+1}{2}\right]^{th}$ item

(d) None of these

46. Which one of the following average can be computed with the help of ogives?

(a) Simple Mean

(b) Weighted Mean

(c) Mode

(d) Median

47. In case of an even number of observations, which of the following is median?

(a) Weighted average of the two middle values

(b) Simple average of the two middle values

(c) Any of the two middle-most values

(d) None of these

48. To calculate __________, it is essential to make class-intervals equal and frequencies have to be

adjusted.

(a) Mean

(b) Mode

(c) Median

(d) None of these

49. is called a positional measure.

(a) Median

(b) Mode

(c) Mean

(d) None of these

50. For ordering shoes of various sizes for resale, a size will be more appropriate.

(a) Modal

(b) Mean

(c) Median

(d) None of these

Answers

Multiple Choice Questions

1. (a)	2. (c)	3. (d)	4. (a)	5. (c)	6. (c)
7. (c)	8. (d)	9. (c)	10. (d)	11. (c)	12. (b)
13. (d)	14. (b)	15. (a)	16. (a)	17. (d)	18. (a)
19. (a)	20. (d)	21. (a)	22. (c)	23. (b)	24. (a)
25. (c)	26. (c)	27. (a)	28. (d)	29. (b)	30. (c)
31. (d)	32. (a)	33. (b)	34. (c)	35. (b)	36. (b)
37. (c)	38. (b)	39. (c)	40. (b)	41. (d)	42. (b)

43. (a) **44. (d)** **45. (b)** **46. (d)** **47. (b)** **48. (b)**

49. (a) **50. (a)**

CHAPTER 11

MEASURES OF DISPERSION

1. CONCEPT AND DEFINITION OF DISPERSION

In the last two chapters, we studied Averages or Measures of Central Tendency. Average indicates representative value of the series around which other values of the series tend to converge. So that average represents the series as a whole. One may now be keen to know how far the various values of the series tend to disperse from each other, or from their average. This brings us to yet another important branch of statistical methods, viz., Measures of Dispersion. Only when we study dispersion along with average of a series that we can have a comprehensive information about the nature and composition of a statistical series. Let us take an illustration to understand the point better. The following data relates to wages paid to 5 workers in three factories, A, B and C.

Factory A	Factory B	Factory C
400	350	50
400	380	75
400	400	400
400	420	725
400	450	750

In all the three factories the arithmetic mean and median is same, i.e., 400. But in factory A, there is no variation between the average wage and the wages paid to the different labourers. In factory ‘B’, there is a small variation between average wage and wages paid to the different workers. But in factory C, there is large variation in the average wage and the wages paid to different workers. The minimum wage is Rs. 50 and maximum wage is Rs. 750. It shows that mean and median do not provide complete information about the composition and character of a series. In order to get a comprehensive picture of the series, we should study measures of dispersion as well.

How is Dispersion of the Series different from Average of the Series?

Average of the series refers to central tendency of the series. It represents behaviour of ail items in the series. But different items tend to differ from each other and from their average. Dispersion measures the extent of this difference. Or, dispersion measures the extent to which different items tend to disperse away from the central tendency.

Definition

According to Dr. Bowley, "Dispersion is the measure of the variation of the items."

In the words of Spiegel, "The degree to which numerical data tend, to spread about an average value is called the variation or dispersion of the data."

Objectives Related to the Measurement of Dispersion

What is the Basic Objective Related to the Measurement of Dispersion?

It **is** to know about the composition of the statistical series by estimating the extent to which different items of the series tend to move away from their average value or the central tendency.

Following are some specific objectives related to the measurement of dispersion:

(i) To know the variation of different values of the items from the average value of a series.

(ii) To know about the composition of a series or the dispersal of values on either sides of the central tendency.

(iii) To know the range of values (i.e., difference between the highest and the lowest value).

(iv) To compare the disparity between two or more series in order to find out the degree of variation.

(v) To know whether the central tendency truly represents the series or not.

Absolute and Relative Measures of Dispersion

There are two measures of dispersion, as discussed under:

(i) Absolute Measure

When dispersion of the series is expressed in terms of the original unit of the series, it is called absolute measure of dispersion. Thus, dispersion of price series would be expressed in terms of rupees; dispersion of weight series would be expressed in terms of kilograms; and so on. For example, if one states that the average wage of a group of the workers is Rs. 100 and dispersion of the wage is Rs. 10, one is referring to absolute dispersion. Absolute measure of dispersion is used when only one set of statistical distribution is under consideration. It cannot be used when comparison is involved across two or more sets of statistical series with different units of measurement (like 'rupee' in one case and 'kilogram' in the other).

(ii) Relative Measure

The relative measure of dispersion expresses the variability of data in terms of some relative value or percentage. Thus, if one states that 26 per cent of the people in India are below poverty line, one is referring to the relative variability of data. In such cases, absolute variability is divided by the mean value of the series or percentage of the absolute variability is determined. This measure of dispersion is used when one studies two or more series simultaneously. Relative measure of dispersion is known as Coefficient of Dispersion.

2. METHODS OF MEASURING DISPERSION

Following are the methods of absolute and relative measures of dispersion:

Absolute Measure	Relative Measure
(1) Range	Coefficient of Range
(2) Quartile Deviation; Inter Quartile Range	Coefficient of Quartile Deviation
(3) Mean Deviation	Coefficient of Mean Deviation
(4) Standard Deviation	Coefficient of Standard Deviation, Coefficient of Variation
(5) Lorenz Curve	

RANGE

It is the simplest method of measuring dispersion of data. Range is the difference between the highest value and the lowest value in a series.

Thus:

FORMULA

R = H - L

Here, R = Range; H = Highest value in the series;

L = Lowest value in the series.

To illustrate, pocket expenses of 5 students are reported to be Rs. 20, Rs. 30, Rs. 40, Rs. 50 and Rs. 100 per month. The highest value (H) in this series is Rs. 100 and the lowest value (L) is Rs. 20. Accordingly,

R = H - L

R = Rs. 100 - Rs. 20 = Rs.80

Thus, R = Rs. SO.

Why should we measure Dispersion about some particular value?

All measures of dispersion do riot measure variation about some particular value of the series. Range for example, is simply the difference between the highest and the lowest values of a statistical series, I ikewise, quartile deviation is defined as half of the difference between Q3 (third quartile) and Q1 (first quartile). But measures of dispersion like mean deviation and standard deviation are worked out as deviations from some central tendency of the statistical series. Deviations from some central tendency (or central value like mean or median) of the series offers a better picture of dispersion. Why? Because:

(i) then we can assess how precise is the central tendency as the representative value of all the observations in the series. Greater value of dispersion implies lesser representativeness of the central tendency and vice versa, and

(ii) we can precisely assess how scattered are the actual observations from their representative value. Actual observations may show both a positive variation as well as negative variation from the observed central tendency of the series. A summary measure of variation (which is what a measure of dispersion focuses on) is possible only when all the variations are duly considered.

Coefficient of Range

Range is an absolute measure of dispersion. As such it cannot be used for comparisons. To make it comparable we find its coefficient. It is the ratio between (i) the difference between the highest and lowest values of the series and (ii) the sum of the lowest and highest values of the series. It is calculated as under:

FORMULA

Coefficient of Range, CR = $\frac{H-L}{H+L}$

(Here, CR — Coefficient of range; H = Highest value in the series; L = Lowest value in the series.)

Thus of the given illustration,

Coefficient of Range, (CR) = $\frac{100-20}{100+20} = \frac{80}{120} = 0.67$

Calculation of Range and Coefficient of Range for Different Types of Statistical Series

Let us understand through different illustrations how range and its coefficient are calculated for different types of statistical series.

(1) Individual Series and Range

In the individual series, range is calculated as the difference between the highest and lowest value of the series.

Illustration.

Monthly wages of workers of a factory are stated below. Find out the range and the coefficient of range.

Wages (Rs.)	50	60	80	90	200	225	250	300	340	360	400	415	425	450	500

Solution:

Range (R) = H - L

Here, H = 500; L - 50

Thus, R = 500 - 50 - 450 ,

Coefficient of Range (CR) = $\frac{H-L}{H+L}$

$$= \frac{500-50}{500+50} = \frac{450}{550} = 0.82$$

Range = 450

Coefficient of Range = 0.82.

(2) Discrete Frequency Series or Frequency Array and Range

As in the case of individual series, range of the discrete series is determined as the difference between the highest value and the lowest value of the series. Frequency of the series is not taken into account.

Illustration.

Calculate range and coefficient of range of the following series.

Size	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**18**
Frequency	**1**	**13**	**24**	**14**	**15**	**13**	**16**	**20**

Note

That in the estimation of range or coefficient of range, frequency of the items is not taken into consideration.

Solution:

Here, H = 18; L = 10

Range (R) = H - L = 18-10 = 8

Coefficient of Range (CR) = $\frac{H-L}{H+L} = \frac{18-10}{18+10} = \frac{8}{28} = 0.29$

Range = 8

Coefficient of Range = 0.29.

(3) Frequency Distribution Series and Range

In case of frequency distribution series, we find the difference between lower limit of the first-class interval and upper limit of the last class interval in the series. Difference between these values would be the range of the series. That is,

FORMULA

R = Upper Limit of the Last Class Interval - Lower Limit of the First-Class Interval

Illustration.

Find out the range and the coefficient of range of the following series:

Marks	Number of Students
20-29	8
30-39	12
40-49	20
50-59	7
60-69	3

Solution:

This is an inclusive series. For the estimation of range, this must be converted into exclusive series, as below:

Marks	Number of Students
19.5-29.5	8
29.5-39.5	12
39.5-49.5	20
49.5-59.5	7
59.5—69.5	3

L =19.5; H = 69.5 Range (R) = 69.5 - 19.5 = 50

Coefficient of Range (CR) = $\frac{69.5-19.5}{69.5+19.5}$

$$= \frac{50}{89}$$

Range = 50

Coefficient of Range = 0.562.

Merits and Demerits of Range as a Measure of Dispersion

Merits

(1) Simple: It is a very simple measure of the dispersion of the series. It is simple to calculate as well as understand.

(2) Widely Used: Range is widely used in statistical series relating to quality control in production. Control charts are prepared on the basis of range. If the quality of goods produced is within the range prescribed in the charts then the production process is said to be under control, otherwise not. Likewise, range is a commonly used measure of dispersion in case of changes in interest rates, exchange rates and share prices.

Demerits

(1) Unstable: It is an unstable measure of dispersion. It depends upon the extreme values of the series. Any change in the extreme values or a change in the sample immediately affects the range of the series.

(2) Not Based on all Values: The calculation of range is not based on all the values of a series. It does not give importance to other values other than the extreme ones.

(3) **No Knowledge of the Formation of the Series:** Range gives no precise knowledge about the formation of series.

(4) Irrelevant for Open-ended Frequency Distribution: Range cannot be calculated in case of open-ended frequency distributions. It just becomes irrelevant.

Inter Quartile Range and Quartile Deviation (QD) and Their Coefficient

Inter Quartile Range

Difference between third Quartile (Q_3) and first Quartile (Q_1) of a series, is called Inter Quartile Range.

FORMULA

Inter Quartile Range = $Q_3 - Q_1$

Quartile Deviation

Quartile Deviation is Half of Inter Quartile Range.

FORMULA

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

It is also called Semi-inter Quartile Range.

Coefficient of Quartile Deviation

Coefficient of quartile deviation is calculated using the following formula:

FORMULA

Coefficient of $QD = \frac{Q_3 - Q_1}{2} \div \frac{Q_3 + Q_1}{2}$

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

Coefficient of Quartile Deviation = $\frac{Q_1 - Q_1}{Q_3 + Q_1}$

Calculation of Quartile Deviation and Coefficient of Quartile Deviation for Different Types of Statistical Series

Through different illustrations, let us understand how quartile deviation and coefficient of quartile deviation are calculated for different statistical series.

(1) Individual Series and Quartile Deviation

In order to calculate quartile deviation in case of individual series, we first find out the values of third quartile and first quartile using the equations given on next page:

Q_1 = Size off $\left(\frac{N+1}{4}\right)$th item

Q_3 = Size of 3$\left(\frac{N+1}{4}\right)$th item

Quartile Deviation (QD) and the coefficient of QD are then calculated using the following formulae:

FORMULAE

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

Coefficient of QD = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

Illustration.

Find out the quartile deviation and coefficient of quartile deviation of the following series:

S. No.	1	2	3	4	5	6	7	8	9	10	11
Marks	10	15	20	25	30	35	40	45	50	55	60

Solution:

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

$$= \text{Size of} \left(\frac{11+1}{4}\right) \text{th item}$$

= Size of 3rd item

= 20 marks

$$Q_3 = \text{ Size of } 3\left(\frac{N+1}{4}\right) \text{ th item}$$

$$= \text{ Size of } 3\left(\frac{11+1}{4}\right) \text{ th item}$$

= Size of 9th item

= 50 marks

Quartile Deviation $(QD) = \frac{Q_3-Q_1}{2} = \frac{50-20}{2} = 15$

Coefficient of Quartile Deviation = $= \frac{Q_3-Q_1}{Q_3+Q_1} = \frac{50-20}{50+20} = \frac{30}{70}$

QD = 15, Coefficient of QD = 0.43

Note

$(Q_3$-$Q_1)$/2 or quartile deviation is average of the difference between two quartiles (Q_3 and Q_1).

Illustration.

The following table shows monthly wages of 10 workers:

Monthly Wages (Rs.)	120	150	170	180	181	187	190	192	200	210

Calculate first, third quartiles and quartile deviation.

Solution:

Calculation of the Partition Values

S. No.	1	2	3	4	5	6	7	8	9	10
Monthly Wages (Rs.)	120	150	170	180	181	187	190	192	200	210

Q_1 = Size of $\left(\frac{N+1}{4}\right)$ th item

= Size of $\left(\frac{10+1}{4}\right)$ th item

= Size of 2.75th item

= Size of 2nd item+ $\frac{3}{4}$(Size of 3rd item −Size of 2nd item)

$$= \text{Rs.}150 + \frac{3}{4}(170 - 150)$$

$Q_1 = \text{Rs.}165$

$Q_3 = \text{Size of } 3\left(\frac{N+1}{4}\right)\text{th item} = \text{Size of } 8.25\text{th item}$

$= \text{Size of 8th item} + \frac{1}{4}$ (Size of 9th item- Size of 8th item)

$Q_3 = \text{Rs.}192 + \frac{1}{4}(200 - 192) = \text{Rs.}194$

$QD = \frac{Q_3 - Q_1}{2} = \frac{194-165}{2} = \frac{29}{2} = 14.5$

(2) Discrete Series or Frequency Array and Quartile Deviation

In a discrete series, quartile deviation is calculated by converting simple frequencies of series into cumulative frequencies. It is illustrated on next page.

Illustration.

The following data shows daily wages of 199 workers of a factory. Find out quartile deviation and the coefficient of quartile deviation.

Wages (Rs.)	10	20	30	40	50	60	70	80	90	100
Number of Workers	2	8	20	35	42	20	28	26	16	9

Solution:

The above series is first converted into a cumulative frequency distribution series.

Wages (Rs.)	**Frequency (f)**	**Cumulative Frequency**
10	2	2
20	8	10
30	20	30
(Q_1) 40	35	**65**
50	42	107

60	20	127
(Q_3) 70	28	155
80	26	181
90	16	197
100	2	199
	N - 199	

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

$$= \text{ Size of } \left(\frac{199+1}{4}\right) \text{ th item = Size of 50th item}$$

50th item lies in 65th cumulative frequency of the series. Wage corresponding to 65th cumulative frequency is Rs. 40 which therefore is first quartile of the wage distribution. Likewise,

$$Q_3 = \text{ Size of } 3\left(\frac{N+1}{4}\right) \text{ th item}$$

$$= \text{ Size of } 3\left(\frac{199+1}{4}\right) \text{ th item = Size of 150th item}$$

150th item falls in 155th cumulative frequency of the series. Wage corresponding to 155th cumulative frequency is Rs. 70 which therefore is the third quartile of the series.

Quartile Deviation $(QD) = \frac{Q_3 - Q_1}{2}$

$$= \frac{70-40}{2} = \frac{30}{2} = 15$$

Coefficient of QD $QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{70-40}{70+40} = \frac{30}{110} = 0.27$

(3) Frequency Distribution Series and Quartile Deviation

Following illustration should explain the calculation of quartile deviation in frequency distribution series:

Illustration.

Find out quartile deviation of the following series:

Age (Years)	0-20	20-40	40-60	60-80	80-100

Number of Persons	4	10	15	20	11

Solution:

Age **(Years)**	**Number of Persons** **(f)**	**Cumulative Frequency**
0-20	4	4
20-40	10	14
40-60	15	29
60-80	20	49
80-100	11	60

$$Q_1 = \text{ Size of } \left(\frac{N}{4}\right) \text{ th item } = \text{ Size of } \left(\frac{60}{4}\right) \text{th item}$$

$$= \text{ Size of 15 th item}$$

15th item lies in group 40-60 and falls within 29th cumulative frequency of the series.

$$Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f} \times i$$

(Here, l_1 = Lower limit of the class interval; N = Sum total of the frequencies; c.f. = Cumulative frequency of the class preceding the first quartile class; f = Frequency of the quartile class; i = Class interval.)

Thus,

$$\text{Q1} = 40 + \frac{\frac{60}{4} - 14}{15} \times 20$$

$$= 40 + \frac{15 - 14}{15} \times 20$$

$$= 40 + \frac{1}{15} \times 20$$

$$= 40 + \ 1.33 \ = 41.33$$

Likewise,

$$Q_3 = \text{Size of } 3\left(\frac{N}{4}\right) \text{th item}$$

$$= \text{Size of } 3\left(\frac{60}{4}\right) \text{th item}$$

$$= \text{Size of 45th item}$$

45th item falls within 49th cumulative frequency of the series.

$$\text{Thus, } Q_3 = i_1 + \frac{3\left(\frac{N}{4}\right) - c.f}{f} \times i$$

$$= 60 + \frac{3\left(\frac{60}{4}\right) - 29}{20} \times 20$$

$$= 60 + \frac{45 - 29}{20} \times 20$$

$$= 60 + \frac{16}{20} \times 20$$

$$= 60 + 16 = 76$$

Having known the values of Ql and Q_3, quartile deviation (QD) is found as,

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

$$= \frac{76 - 41.33}{2} = \frac{34.67}{2} = 17.34$$

And,

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_I}$$

$$= \frac{76 - 41..33}{76 + 4133} = \frac{34.67}{117.33}$$

$$= 0.30$$

Thus,

QD = 17.34, and Coefficient of QD = 0.30.

Merits and Demerits of Quartile Deviation

Merits

(1) Simple: It is very simple to calculate and understand.

(2) Less Effect of Extreme Values: Quartile deviation is less affected by extreme values of the series.

Demerits

(1) Not Based on all Values: The calculation of quartile deviation is not based on all values of the series. It is, therefore, less representative.

(2) Formation of Series not known: this method does not show complete formation of the series.

(3) Instability: The calculation of quartile deviation is significantly influenced by change in sample of the population. Accordingly, it suffers from instability.

MEAN DEVIATION

Mean Deviation is the arithmetic average of the deviations of all the values taken from some average value (mean, median, mode) of the series, ignoring signs (+ or -) of the deviations.

In the words of Clark and Schakde, "Mean Deviation is the Arithmetic Average of deviations of all the values taken from a statistical average (mean, median, or mode) of series. In taking deviation of values, algebraic signs + and - are not taken into consideration, that is negative deviations are also treated as positive deviations."

Calculation of mean deviation involves the following steps:

(i) We first of all find out mean, median or mode of a series.

(ii) As a second step, we find out the deviations of different items from the (central value mean, median or mode of the series.)

These deviations are added up. While adding up these deviations positive (+) and negative (-) signs are ignored. All deviations are treated as positive.

(iii) On both sides of the deviation, from the mean, are drawn two straight lines signifying that while calculating deviation negative signs have been ignored and all deviations have been treated as positive.

(iv) Mean deviation is known by dividing the sum total of the deviation by the number of items.

Note: Deviations are often found from median value of the series. Many a time mean is also used, bur the use of Modal value is seldom made. Formula of mean deviation is as follows:

FORMULA

If deviations are taken from median, the following formula is used:

$$MD_m = \frac{\Sigma|X - M|}{N} \text{ or } \frac{\Sigma|dm|}{N}$$

And, if deviations are taken from arithmetic average of the series, then

$$MD_{\bar{X}} = \frac{\Sigma|X - \bar{X}|}{N} \text{ or } \frac{\Sigma|d\bar{x}|}{N}$$

(Here, MD= Mean deviation; X-M = Deviation from the median; X - X = Deviation from the arithmetic average; N = Number of items.)

Coefficient of Mean Deviation

In order to find out coefficient of mean deviation, mean deviation of the series is divided by the central tendency of the series. If the deviations are taken from arithmetic mean, the mean deviation is divided by the arithmetic mean. And, if the deviations are taken from median, the mean deviation is divided by the median. Likewise, if the deviations are taken from mode of the series, the mean deviation is divided by the mode value.

Thus,

(1) Coefficient of MD from Mean $= \frac{MD_{\bar{x}}}{\bar{X}} = \frac{\text{Mean Deviation}}{\text{Arithmetic Mean}}$

(2) Coefficient of MD from Median $= \frac{MD_m}{M} = \frac{\text{Mean Deviation}}{\text{Median}}$

(3) Coefficient of MD from Mode $= \frac{MD_2}{Z} = \frac{\text{Mean Deviation}}{\text{Mode}}$

Calculation of Mean Deviation

(1) Individual Series and Mean Deviation

Following illustration should explain the calculation of mean deviation in case of individual series:

Illustration.

The data below gives wages of workers in a factory. Find out mean deviation and its coefficient.

S. No.	1	2	3	4	5	6	7	8	9
Wages (Rs.)	40	42	45	47	50	51	54	55	57

Solution:

Steps: Estimating MD through Median	**Steps: Estimating MD through Mean**
(i) Arrange the data in ascending order.	(i) Calculate arithmetic mean by adding up the data.

Calculate the median $M = \text{Size of } \left(\frac{N+1}{2}\right) \text{ th item}$	(ii) Find out the deviation of the items from mean. Ignore (+) and (-) signs. Express it by | dx |.
Find out the deviation of the items from median. Ignore (-) and (+) signs. Express it by |dm| sign.	(iii) Sum up the deviations.
Sum up the deviations, and express the same by using the relevant formula to find the required answer. $MD_m = \frac{\Sigma\|dm\|}{N}$	(iv) Use the relevant formula to find the required answer. $MD_{\bar{x}} = \frac{\Sigma\|d\bar{x}\|}{N}$
After that solve the question with the help of the formula.	

Calculation of Mean Deviation and Coefficient of Mean Deviation using Median			Calculation of Mean Deviation and Coefficient of Mean Deviation using Arithmetic Mean		
S. No.	**Wages** (Rs.)	**Deviation from Median** **|dm| = |X-M| M = 50**	**S. No.**	**Wages** (Rs.)	**Deviation from Arithmetic Mean** $\|d\bar{x}\| = \|X - \bar{x}\|$ $\bar{X}$ **=49**
1	40	10	1	40	9
2	42	8	9	42	7
3	45	5	3	45	4
4	47	3	4	47	2
5	50 (M)	0	5	50	**1**
6	51	1	6	51	2

7	54	4	7	54	5
8	55	5	8	55	6
9	57	7	9	57	8
N=9		Σ | dm | = 43	N= 9	∑X = 441	Σ | dx | = 44

From Median	From Mean
(a) M = Size of $\left(\frac{N+1}{2}\right)$th item M = Size of $\left(\frac{9+1}{2}\right)$th item M = 5th item = Rs. 50 (b) $\text{MD}_\text{m} = \frac{\Sigma\lvert\text{dm}\rvert}{\text{N}} = \frac{43}{9} = 74.78$ (c) Coefficient of MD_m $= \frac{\text{MD}_\text{m}}{\text{M}} = \frac{4.78}{50} = 0.096$	$\overline{\text{X}} = \frac{\Sigma\text{X}}{\text{N}}$ $= \frac{441}{9}$ = Rs.49 (b) $MD\bar{\text{x}} = \frac{\Sigma\lvert\text{d}\bar{\text{x}}}{\text{N}} = \frac{44}{9} = \bar{z}4.89$ (c) Coefficient of $\text{MD}_{\bar{\text{x}}} = \frac{\text{MD}_{\overline{\text{X}}}}{\overline{\text{X}}} = \frac{4.89}{49}$

(2) Discrete Series or Frequency Array and Mean Deviation

Following steps may be noted in the calculation of mean deviation for the discrete series:

(i) Find out central tendency of the series (mean or median) from which deviations are to be taken.

(ii) Deviations of different items in the series are taken from central tendency, and signs (+) or (-) of the deviations are ignored. It is expressed as (| dx |) or (| dm |).

(iii) Each deviation value is multiplied by the frequency facing it, and the sum of these multiples is obtained. This is expressed as ∑f | d |.

(iv) ∑f | d | is divided by the sum total of frequencies, that is ∑f or N. The resultant value would be mean deviation. ,

Thus,

$$MD_m = \frac{\Sigma f|dm|}{N}$$

[Here, ‘dm’ indicates that deviations are taken from median (M) of the series.]

Illustration.

Using median and arithmetic mean respectively, calculate mean deviation and its coefficient from the following data:

Size of Items	5	6	7	8	9	10	11	12	13
Frequency	4	5	6	7	8	9	10	11	12

Solution:

(i) Calculation of Mean Deviation from Median

Size of Items (X)	**Frequency** (f)	**Cumulative** **Frequency**	**Deviation from Median** **| dm | = | X - M |** **M=10**	**Multiple of Deviation and the Corresponding Frequency** **f|dm|**
5	4	4	5	20
6	5	9	4	20
7	6	15	3	18
8	7	22	2	14
9	8	30	1	8
10	9	39	0	0
11	10	49	1	10
12	11	60	2	22
13	12	72	3	36

	N = 72			∑f|dm| = 148

(a) Median or M = Size of $\left(\frac{N+1}{2}\right)$th item

= Size of $\left(\frac{172+1}{2}\right)$th item

= Size of 36.5th item = 10

(b) Mean Deviation from Median

Or $\mathrm{MD_m} = \frac{\sum f|dm|}{N} = \frac{148}{72}$

= 2.05

(c) Coefficient of Mean Deviation from Median

$$= \frac{\mathrm{MD_m}}{\mathrm{M}} = \frac{2.05}{10}$$

= 0.205

(ii) Calculation of Mean Deviation from Arithmetic Mean

Size of Items (X)	**Frequency (f)**	**fX**	**Deviation from Mean (9.83) (-), (+) ignored** $|d\bar{x}|$	**Product of Deviation and the Corresponding Frequency (f** $|d\bar{x}|$**)**
5	4	20	4.83	19.32
6	5	30	3.83	19.15
7	6	42	2.83	16.98
8	7	56	1.83	12.81
9	8	72	0.83	6.64
10	9	90	0.17	1.53
11	10	110	1.17	11.70
12	11	132	2.17	23.87

13	12	156	3.17	38.04
	∑f = 72	**∑fX = 708**		**∑F** $\vert d\bar{x}\vert$**=150.04**

(a) $\overline{X} = \frac{\Sigma fX}{\Sigma f} = \frac{708}{72} = 9.83$

(b) Mean Deviation from Arithmetic Mean

Or, $MD_{\overline{\lambda}} = \frac{\Sigma f|d\overline{x}|}{\Sigma f} = \frac{150.04}{72} = \ 2.08$

(c) Coefficient of Mean Deviation from Mean

$$= \frac{MD_{\overline{x}}}{\overline{X}} = \frac{2.08}{9.83} = \ 0.21$$

$$= 0.21$$

(3) Frequency Distribution Series and Mean Deviation

Continuous series are first converted into discrete series by finding mid-values of the class intervals. Afterwards, same procedure is followed for the calculation of mean deviation and its coefficient as in the case of discrete series.

Illustration.

Find out mean deviation and coefficient of mean deviation, using arithmetic mean from the following data:

Profit (Rs.)	0-10	10-20	20-30	30-40	40-50
Shops (Number)	5	10	15	20	25

Solution:

Profit **(Rs.)**	**Mid-value** **(m)**	**Frequency** **(f)**	**Multiple of Mid-value and Frequency (fm)**	**Deviation from Mean** $\vert d\bar{x}\vert = \vert m - \bar{x}\vert$ **X=31.66**	**Multiple of Deviation and the Corresponding to Frequency (f**$\vert d\bar{x}\vert$**)**
0-10	5	5	25	26.66	133.30
10-20	15	10	150	16.66	166.60

20-30	25	15	375	6.66	99.90
30-40	35	20	700	3.34	66.80
40-50	45	25	1,125	13.34	333.50
		∑f = 75	**∑fm = 2,375**	$\sum\|d\bar{x}\|$ = 66.66	**∑f**$\|d\bar{x}\|$ **= 800.10**

$$\bar{X} = \frac{\sum fm}{\sum f} = \frac{2,375}{75} = 31.66$$

Mean Deviation from Arithmetic Mean

$$MD_{\bar{X}} = \frac{\Sigma f|dx|}{\Sigma f} = \frac{800 \cdot 10}{75} = 10.67$$

Coefficient of Mean Deviation

$$= \frac{MD_{\bar{X}}}{\bar{X}} = \frac{10.67}{31.67} = 0.34$$

Mean Deviation = 10.67, Coefficient of Mean Deviation = 0.34.

Illustration.

Calculate mean deviation and its coefficient from the median of the following data:

Size	100-120	120-140	140-160	160-180	180-200
Frequency	4	6	10	8	5

Solution:

Size **(X)**	**Mid-value** **(m)**	**Frequency** **(f)**	**Cumulative** **Frequency**	**Deviation from Median** **| dm | = | m - M | M = 153**	**Multiple of Deviation and the Corresponding Frequency (f|dm|)**
100-120	110	4	4	110-153 = 43	172
120-140	130	6	10 (c.f.)	130-153 = 23	138
140-160	150	10(f)	20	150-153 = 3	30

160-180	170	8	28	170-153 = 17	136
180-200	190	5	33	190 -153 = 37	185
		∑f = 33			**∑f | dm | = 661**

$\mathrm{M} =$ Size of $\left(\frac{\mathrm{N}}{2}\right)$ th item = Size of $\left(\frac{33}{2}\right)$ th item

= Size of 16.5 th item

140-160 is the Median Class Interval.

The Median is calculated as

$$\mathrm{M} = l_1 + \frac{\frac{\mathrm{N}}{2} - \mathrm{c.f}}{\mathrm{f}} \times l$$

$$= 140 + \frac{16.\tilde{o} - 10}{10} \times 20$$

$$= 140 + 13 = 153$$

Mean Deviation from Median

$$\mathrm{MD_m} = \frac{\Sigma f|dm|}{\Sigma f} = \frac{661}{33} = 20.03$$

Coefficient of Mean Deviation

$$= \frac{\mathrm{MD_m}}{\mathrm{M}} = \frac{20.03}{153} = 0.\ 1309$$

$\mathrm{MD_{n]}} = 20.03$, Cocfricient of $\mathrm{MD_m} = 0.1309$.

Merits and Demerits of Mean Deviation

What is the Principal Drawback of Mean Deviation as a Measure of Dispersion?

It is that all deviations from the average value of the series are taken as positive, even when some of these are actually negative.

Merits

(1) Simple: It is a very simple and easy measure of dispersion.

(2) Based on all Values: Mean deviation is based on all the items of the series. It is therefore more representative than the range or quartile deviation.

(3) Less Effect of Extreme Values: Mean deviation is less affected by extreme values than the range.

Demerits

(1) Inaccuracy: Calculation of mean deviation suffers from inaccuracy, because the *+’ or signs are ignored.

(2) Not Capable of Algebraic Treatment: Mean deviation is not capable of any further algebraic treatment.

(3) Unreliable: In case deviations are taken from mode and mode being uncertain, mean deviation also becomes uncertain and therefore, unreliable.

STANDARD DEVIATION

Standard deviation is a most satisfactory scientific method of dispersion. Accordingly, it is a widely used method in statistical analysis.

This was first used by **Karl Pearson.** This is sometimes called as ‘Root Mean Square Deviation’. This is generally denoted by {sigma) of the Greek language. Standard Deviation is the square root of the arithmetic mean of the squares of deviations of the items from their mean value. Standard deviation has two main features: (i) The value of its deviation is taken from arithmetic mean, (ii) Plus and minus signs of the deviations taken from the mean are not ignored. In fact, signs of the deviations become redundant once the deviations are squared. Finally, square root of the arithmetic mean of the squares of the deviation is calculated. It is this square root which is called Standard Deviation. This is always in positive value.

In the words of **Spiegel,** “The Standard deviation is the square root of the arithmetic mean of the squares if all deviations. Deviations being measured from arithmetic mean of the items.”

Coefficient of Standard Deviation

This is a relative measure of the dispersion of series. It is generally used whenever variation in different series is compared.

Coefficient of standard deviation is estimated by dividing the value of standard deviation by the mean of the series. Thus,

FORMULA

Coefficient of Standard Deviation = $\frac{\sigma}{\bar{X}}$

Calculation of Standard Deviation

(1) Individual Series and Standard Deviation

There are three methods of calculating standard deviation in case of individual series:

(i) Direct Method,

(ii) Short-cut Method, and

(iii) Step-deviation Method.

(i) Direct Method

Direct method of calculating standard deviation is most useful when mean value is in whole number. This method involves the following steps:

(a) First of all, mean value of the concerned series is determined. That is, we find out $\bar{X}$ as, $\bar{X} = \frac{\Sigma X}{N}$

(b) Deviation of each item from X is determined. That is, we find the values of x as, x = X - X

(c) Each value of the deviation is squared. The sum total of the square of the deviations is obtained. That is, we find out $\sum x^2$.

(d) $\sum x^2$ is divided by the number of items (N) in the series. Square root of $\frac{\sum x^2}{N}$ will be the standard deviation. That is, we calculate the value of $\sqrt{\frac{\Sigma x^2}{N}}$. Thus, following is the formula for the calculation of standard deviation:

FORMULA

$$\text{SD or } \sigma = \sqrt{\frac{\sum x^2}{N}} \text{ or } \sqrt{\frac{\sum (x - \bar{x})^2}{N}}$$

(Here, σ = Standard deviation; $\sum x^2$ = Sum total of the squares of deviations; X = Mean value; X - X = Deviation from mean value; N = Number of items.)'

How Standard Deviation differs from Mean Deviation?

Standard deviation is a most satisfactory scientific method of dispersion. Accordingly, it is a widely used method in statistical analysis.

Two main differences between standard deviation and mean deviation are:

(i) In the calculation of standard deviation, deviations are taken only from the mean value of the series. On the other hand, in the calculation of mean deviation, deviations may be taken from mean, median or mode.

(ii) In the calculation of mean deviation, signs of deviations (+) or (-) are ignored. But in the calculation of standard deviation, signs are not ignored.

Illustration.

Following are the marks obtained by 10 students of a class. Calculate standard deviation and coefficient of standard deviation.

Marks	12	8	17	13	15	9	18	11	6	1

Solution:

Calculation of Standard Deviation Using Direct Method

S. No.	**Marks** **(X)**	**Deviation** **(x = X - $\bar{X}$) X = 11**	**Square of Deviation** **x^2= (X-$\bar{X}$)^2**
1	12	12- 11 = 1	1
2	8	8-11=-3	9
3	17	17 - 11 = 6	36
4	13	13-11 = 2	4
5	15	15- 11 = 4	16
6	9	9-11 = - 2	4
7	18	18-11 = 7	49
8	11	11 - 11 = 0	0
9	6	6-11=-5	25
10	1	1-11 =-10	100
N = 10	**∑X=110**	**∑X=0**	**$\sum x^2$ = 244**

$$\bar{X} = \frac{\sum X}{N} = \frac{110}{10} = 11$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{244}{10}} = \sqrt{24.4} = 4.94$$

Coefficient of SD= $\frac{\sigma}{\bar{X}} = \frac{4.94}{11} = 0.45$

SD = 4.94 marks, Coefficient of SD = 0.45.

(ii) Short-cut Method

Short-cut method of calculating standard deviation involves the following steps:

(a) We take any value of the series as assumed average, generally written as A.

(b) Deviations of all the items are obtained from the assumed average. Sum total of these deviations is obtained as Σ(X - A) or ∑dx.

Also, we square up the deviations and obtain their sum total as $(X - A)^2$ or $\sum dx^2$.

(c) The following formula is applied to calculate the value of standard deviation:

FORMULA

$$\sigma = \sqrt{\frac{\sum dx^2}{N} - \left(\frac{\sum dx}{N}\right)^2}$$

Illustration.

Find out standard deviation, given the following data: 8, 10, 12, 14, 16, 18, 20, 22, 24, 26

Solution:

S. No.	**Size** **(X)**	**Deviation from Assumed Average** **(dx = X - A)** **A = 20**	**Square of Deviation** **(dx^2)**
1	8	8-20 = - 12	144
2	10	10-20 = - 10	100
3	12	12-20 = -8	64
4	14	14-20 = -6	36
5	16	16-20 = -4	16
6	18	18-20 =-2	4
7	20 (A)	20 - 20 = 0	0
8	22	22 - 20 = + 2	4
9	24	24 - 20 =+ 4	16

10	26	26-20 = + 6	36
N = 10		**∑dx =- 30**	**$\sum dx^2$ = 420**

$$\sigma = \sqrt{\frac{\sum dx^2}{N} - (\frac{\sum dx}{N})^2}$$

$$= \sqrt{\frac{420}{10} - (\frac{-30}{10})^2}$$

$$= \sqrt{42 - (-3)^2} = \sqrt{33} = 5.74$$

Standard Deviation = 5.74.

(iii) Step-deviation Method

This method involves the following steps:

(a) We take any value of the series as assumed average.

(b) Deviations are taken from the assumed average as dx = (X-A).

(c) The deviations are divided by some common factor, as $dx' = \frac{dx}{C}$ = where **C** is the common factor and dx' are step- deviations.

(d) Sum of the step-deviations is obtained. Also, step-deviations are squared and then sum total is obtained as $\sum dx^2$.

(e) The following formula is applied to calculate the value of standard deviation.

FORMULA

$$\sigma = \sqrt{\frac{\sum dx'^2}{N} - (\frac{\sum dx'}{N})^2} \times c$$

Illustration.

Find out standard deviation of the monthly income of 5 persons, as stated below:

S. No. of Persons	Monthly Income (in Rs.)
1	500
2	700
3	1.000

4	1,500
5	1,300

Solution:

S. No.	Monthly Income	Deviation from Assumed Average (dx = X - A) A = 1,000	$dx' = \frac{dx}{C}$ C = 100	dx'^2
1	500	-500	-5	25
2	700	-300	-3	9
3	1.000(A)	0	0	0
4	1,500	+ 500	+ 5	25
5	1,300	+ 300	+ 3	9
N = 5			**∑dx'=0**	**∑dx'²=68**

$$\sigma = \sqrt{\frac{\Sigma DX'^2}{N} - \left(\frac{\Sigma DX'}{N}\right)^2} \times C$$

$$= \sqrt{\frac{68}{5} - \left(\frac{0}{5}\right)^2} \times 100$$

$$= \sqrt{13.6} \times 100$$

= 3.6878 × 100 = 368.78

Standard Deviation (σ) = 368.78.

(iv) Yet another Method

Standard deviation in case of individual series can be calculated using the following formula:

Illustration.

$$\sigma = \sqrt{\frac{\sum x^2}{N} - (\frac{\sum x}{N})^2}$$

Using electronic calculator, find out standard deviation of the following data:

Marks	10	20	30	40	50

Solution:

Calculation of Standard Deviation

S. No.	Marks (X)	X^2
1	10	100
2	20	400
3	30	900
4	40	1,600
5	50	2,500
N = 5	**∑X = 150**	**$\sum X^2$ = 5,500**

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

$$= \sqrt{\frac{5,500}{5} - (\frac{150}{5})^2} = \sqrt{1,100 - (30)^2}$$

$$= \sqrt{1,100 - 900} = \sqrt{200} = 14.14$$

Standard Deviation (σ) = 14.14.

(2) Discrete Series or Frequency Array and Standard Deviation

There are two methods of calculating standard deviation in discrete series:

(i) Direct Method, and

(ii) Short-cut Method.

(i) Direct Method

This method involves the following steps:

(a) We first determine mean value of the series as

$$\bar{X} = \frac{\sum fX}{N}$$

(b) Deviations of various items are obtained from the mean value as

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

(c) Squares of deviations are obtained as x^2

(d) Squared deviations are multiplied by their corresponding frequencies, and their sum total is obtained as $\sum fx^2$

(e) The following formula is applied to calculate the value of standard deviation.

FORMULA

$$\sigma = \sqrt{\frac{\sum fx^2}{N}} \text{ or } \sigma = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}}$$

Illustration.

Find out standard deviation of the following data, using direct method:

Size	4	6	8	10	12	14	16
Frequency	1	2	3	5	3	2	**1**

Solution:

Size **(X)**	**Frequency** **(f)**	**Multiple of Size and Frequency (fX)**	**Deviation from Mean** $(x \equiv X - \bar{X})$ **X = 10**	**Square of Deviation** **(x^2)**	**fx^2**
4	1	4	-6	36	36
6	2	12	-4	16	32
8	3	24	‘ -2	4	12
10	5	50	0	0	0

12	3	36	+2	4	12
14	2	28	+4	16	32
16	1	16	+ 6	36	36
	N = 17	$\sum fX = 170$			$\sum fx^2 = 160$

$$\bar{X} = \frac{\sum fX}{N} = \frac{170}{17} = 10$$

$$\sigma = \sqrt{\frac{\sum fx2}{N}}$$

$$= \sqrt{\frac{\sum f(X - \bar{x})^2}{N}}$$

$$= \sqrt{\frac{160}{17}}$$

$$= \sqrt{9.41}$$

$$= 3.07$$

Standard Deviation (σ) = 3.07.

(ii) Short-cut Method

This method involves the following steps:

(a) We take any value of the series as assumed average, written as A. Generally, the value of the item with the highest frequency is taken as assumed average.

(b) Deviations of different items from assumed average are obtained as d = (X - A).

(c) Deviations are multiplied by their corresponding frequencies and then sum total is obtained as ∑fd. Also, deviations are squared and multiplied by the corresponding frequencies to obtain $\sum fd^2$.

(d) The following formula is applied to calculate the value of standard deviation:

FORMULA

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Illustration.

Find out standard deviation of the following data:

Size	1	2	3	4	5	6	7	8
Frequency	5	10	15	20	15	10	10	15

Solution:

Calculation of Standard Deviation (σ) in Discrete Series (Short-cut method)

Size **(X)**	**Frequency** **(f)**	**Deviation from Assumed Average** **(dx = X - A)** **A = 5**	**Square of Deviation** **(dx^2)**	**Multiple of Deviation and the Corresponding Frequency** **(fdx)**	**fdx^2**
1	5	-4	16	-20	80
2	10	-3	9	-30	90
3	15	-2	4	-30	60
4	20	- 1	1	-20	20
5	15	0	0	0	0
6	10	+ 1	1	+ 10	10
7	10	+ 2	4	+20	40
8	15	+ 3	9	+45	135
	∑f = 100			∑fdx = -25	$\sum fdx^2$=435

$$\sigma = \sqrt{\frac{\sum dx^2}{N} - (\frac{\sum dx}{N})^2}$$

$$= \sqrt{\frac{435}{100} - (\frac{-25}{10})^2} = \sqrt{\frac{435}{100} - (\frac{-1}{4})^2}$$

$$= \sqrt{\frac{435}{100} - \frac{1}{16}} = \sqrt{4.29} = 2.07$$

Standard Deviation (σ) = 2.07.

(3) Frequency Distribution Series and Standard Deviation

Three methods are available for the calculation of standard deviation in case of frequency distribution series:

(i) Direct Method,

(ii) Short-cut Method, and

(iii) Step-deviation Method.

(i) Direct Method

This method involves the following steps:

(a) First, mean of the series ($\bar{X}$) is determined.

(b) Deviations of various mid-values are taken from the mean value, x = m - $\bar{X}$ (Here, mid-value is m).

(c) Deviations are squared (x^2) and then multiplied by their corresponding frequencies to get $\sum fx^2$.

(d) Following formula is used to calculate the value of standard deviation:

FORMULA

$$\sigma = \sqrt{\frac{\sum fx^2}{N}} \text{ or } \sigma = \sqrt{\frac{\sum f(X-\bar{X})^2}{N}}$$

Illustration.

Given the following series, calculate standard deviation by direct method:

Size	0-2	2-4	4-6	6-8	8-10	10-12
Frequency	2	4	6	4	2	6

Solution:

Size (X)	Mid-value (m)	Frequency (f)	Multiple of Mid-value and Frequency (fin)	Deviation from Mean Value (x = m -$\bar{X}$) X = 6.5	Square of Deviation (x^2)	fx^2
0-2	1	2	2	-5.5	30.25	60.50
2-4	3	4	12	-3.5	12.25	49.00
4-6	5	6	30	-1.5	2.25	13.50
6-8	7	4	28	+ 0.5	0.25	1.00
8-10	9	2	18	4 2.5	6.25	12.50
10-12	11	6	66	+ 4.5	20.25	121.50
		N = 24	Σfm = 156			$\sum fx^2$=258

$$\bar{X} = \frac{\sum \text{fm}}{\text{N}} = \frac{156}{24} = 6.5$$

$$\sigma = \sqrt{\frac{\sum f x^2}{N}} = \sqrt{\frac{258}{24}}$$

$$= \sqrt{10.75} = 3.28$$

Standard Deviation (σ) = 3.28.

(ii) Short-cut Method

This method is the same as used in case of discrete series. The only difference is that whereas in discrete series, deviations are obtained of the actual values of the series, in case of frequency distribution series deviations are obtained of the mid-values of the class intervals. In both the cases, deviations are taken from the mean value of the series. Thus,

FORMULA

$$\text{SD or } \sigma = \sqrt{\frac{\sum \text{fdx}^2}{\text{N}} - (\frac{\sum \text{fdx}}{\text{N}})^2}$$

Illustration.

Using short-cut method, calculate standard deviation of the following series:

Size	**0-2**	**2-4**	**4-6**	**6-8**	**8-10**	**10-12**
Frequency	**2**	**4**	**6**	**4**	**2**	**6**

Solution:

Size **(X)**	**Mid** **value** **(m)**	**Frequency** **(f)**	**Deviation from Assumed Average** **(dx = m - A)** **A=5**	**Square of Deviation** **(dx^2)**	**Multiple of Deviation and the** **Corresponding** **Frequency** **(fdx)**	**fdx^2**
0-2	1	2	-4	16	-8	32
2-4	3	4	-2	4	-8	16
4-6	5	6	0	0	0	0
6-8	7	4	+ 2	4	4- 8	16
8-10	9	2	+ 4	16	+ 8	32
10-12	11	6	+ 6	36	+ 36	216
		N = 24			**∑fdx=36**	**∑fd x^2=312**

$$\sigma = \sqrt{\frac{\Sigma fdx^2}{N} - (\frac{\Sigma fdx}{N})^2} = \sqrt{\frac{312}{24} - (\frac{36}{24})^2}$$

$$= \sqrt{13 - 2.25} = \sqrt{10.75} = 3.28$$

Standard Deviation (σ) = 3.28.

(iii) Step-deviation Method

This is the most popular method of calculating standard deviation in case of frequency distribution series. It involves the following steps:

(a) Take any value as assumed average, A.

(b) Find out mid-values of the class intervals. Take deviations of the mid-values from A, expressed as ‘dx’.

(c) Divide the deviations by their common factor to get $\left(\frac{dx}{C}\right)$expressed as dx'.

(d) Multiply dx' with the corresponding frequencies and find their sum total as ∑fdx'. Also take squares of (dx'^2), and multiply them by the corresponding frequencies to get ∑fdx'-.

(e) Calculate the value of standard deviation, using the following formula:

FORMULA

$$\sigma = \sqrt{\frac{\sum fdx'^2}{N} \left(\frac{\sum fdx'}{N}\right)^2} \times C$$

Illustration.

Using step-deviation method, calculate standard deviation of the following series:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of Students	5	10	20	40	30	20	10	4

Solution:

Frequency Distribution Series and Standard Deviation (Step-deviation Method)

Marks **(X)**	**Mid-value** **(m)**	**Frequency** **(f)**	**Deviation from Assumed Average** **(dx = m - A)** **A = 35**	$dx' = \frac{dx}{C}$ **c = 10**	**fdx'**	**fdx'²**
0-10	5	5	-30	-3	- 15	45
10-20	15	10	-20	-2	-20	40
20-30	25	20	- 10	- 1	-20	20
30-40	35(A)	40	0	0	0	0
40-50	45	30	+ 10	+ 1	+30	30
50-60	55	20	+ 20	+2	+40	80

60-70	65	10	+ 30	+ 3	+30	90
70-80	75	4	+ 40	+4	+ 16	64
		N = 139			**∑fdx'=61**	**∑fdx'²=369**

$$\sigma = \sqrt{\frac{\Sigma fdx'^2}{N} - (\frac{\Sigma fdx'}{N})^2} \times C$$

$$= \sqrt{\frac{369}{139} - (\frac{61}{139})^2} \times 10$$

$$= \sqrt{2.655 - 0.193}$$

$$= \sqrt{2.462} \times 10$$

$$= 1.569 \times 10 = 15.69$$

Standard Deviation (σ) = 15.69.

Illustration.

Find out standard deviation of the following data-set, using step- deviation method:

Marks	**20-40**	**40-60**	**60-80**	**80-100**	**100-120**	**120-140**
Number of Students	**6**	**9**	**8**	**10**	**11**	**6**

Solution:

Marks **(X)**	**Mid-value** **(m)**	**Frequency** **(f)**	**Deviation from Assumed Average** **(dx = m - A)** **A = 90**	$dx' = \frac{dx}{C}$ **C = 20**	**fdx'**	**fdx'²**
20-40	30	6	-60	-3	- 18	54
40-60	50	9	-40	-2	- 18	36
60-80	70	8	-20	- 1	-8	8

80-100	90	10	0	0	0	0
100-120	110	11	+20	+ 1	11	11
120-140	130	6	+40	+ 2	12	24
		N = 50			**∑fdx' = - 21**	**∑fdx'²= 133**

$$\sigma = \sqrt{\frac{\sum fdx'^2}{N} - (\frac{\sum fdx'}{N})^2} \times C$$

$$= \sqrt{\frac{133}{50} - (\frac{-21}{50})^2} \times 20 \quad [\text{ Here ,C} = 20]$$

$$= \sqrt{2.66 - 0.1764} \times 20$$

$$= \sqrt{2.4836} \times 20$$

$$= 1.576 \times 20 = 31.52$$

Standard Deviation (σ) = 31.52.

Combined Standard Deviation

Just as it is possible to calculate combined mean of two or more groups, similarly the combined standard deviation of two or more groups can be calculated. The combined standard deviation of two groups is denoted by σ_{12} and is computed as follows:

FORMULA

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

Where, σ_{l2} = Combined standard deviation;

σ_l = Standard deviation of the first group;

σ_2 = Standard deviation of the second group;

$d_l = \bar{X}_1 - \bar{X}_{12}, d_2 = \bar{X}_2 - \bar{X}_{12}$

The above formula can be extended to calculate the standard deviation of three or more groups. For example, combined standard deviation of three groups is given by:

FORMULA

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1d_1^2 + N_2d_2^2 + N_3d_3^2}{N_1 + N_2 + N_3}}$$

$$\text{Where, } d_1 = \overline{X}_1 - \overline{X}_{12j}; d_2 = \overline{X}_2 - \overline{X}_{123}; d_3 = \overline{X}_3 - \overline{X}_{123}$$

Illustration.

Two samples of size 100 and 150 respectively have means 50 and 60 and standard deviations 5 and 6. Find the mean and standard deviation of the combined sample of size 250.

Solution:

Given: $N_1 = 100, \quad \bar{x}_1 = 50, \quad \sigma_1 = 5$ $N_2 = 150, \quad \overline{X}_2 = 60, \quad \sigma_2 = 6$

Now,

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

$$= \frac{100 \times 50 + 150 \times 60}{100 + 150} = \frac{5{,}000 + 9{,}000}{250}$$

$$= \frac{14{,}000}{250} = 56$$

$$d_1 = \overline{X}_1 - \overline{X}_{12} = 50 - 56 = -6$$

$$d_2 = \overline{X}_2 - \overline{X}_{12} = 60 - 56 = +4$$

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$$= \sqrt{\frac{100 \times (5)^2 + 150 \times (6)^2 + 100 \times (-6)^2 + 150 \times (4)^2}{100 + 150}}$$

$$= \sqrt{\frac{100 \times 25 + 150 \times 36 + 100 \times 36 + 150 \times 16}{250}}$$

$$= \sqrt{\frac{2{,}500 + 5{,}400 +^{1} 3{,}600 + 2{,}400}{250}} = \sqrt{\frac{13{,}900}{250}}$$

$$= \sqrt{556} = 7.46$$

Hence, the Combined Mean = 56 and the Combined Standard Deviation = 7.46.

Variance

Variance is another measure of dispersion. The term variance was first used by R.A. Fisher in 1918. Variance is the square of the standard deviation. Symbolically,

FORMULA

Variance = $(SD)^2 - \sigma^2$

Calculation of Variance

(i) Variance = $\frac{\Sigma f(X-\bar{X})^2}{N}$ (Actual Mean Method)

(ii) Variance = $\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2$ (Assumed Mean Method)

(iii) Variance = $\left[\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2\right] \times C^2$ (Step-deviation Method)

Illustration.

Calculate the mean and variance from the data given below:

Daily Wages	0-10	10-20	20-30	30-4	40-50
Number of Workers	2	7	10	5	3

Solution:

Daily Wages	**Frequency (f)**	**Mid-value (m)**	**Deviation from Assumed Average (d = m-25) A = 25**	$d' = \frac{d}{C}$ **C = 10**	**fd¹**	**fd'²**
0-10	2	5	-20	- 2	-4	8
10-20	7	15	- 10	- 1	-7	7
20-30	10	25 (A)	0	0	0	0
30-40	5	35	+ 10	+ 1	+ 5	5
40-50	3	45	+ 20	+ 2	+6	12
	∑f = N= 27				**∑fd'=0**	**∑fd'²=32**

$$\bar{X} = A + \frac{\Sigma fd'}{\Sigma f} \times C$$

$$= 25 + \frac{0}{27} \times 10$$

= 25

$$\text{Variance } (\sigma^2) = [\frac{\Sigma fd^2}{N} - (\frac{\Sigma fd'}{N})^2] \times C^2$$

$$= \left[\frac{32}{27} - \left(\frac{0}{27}\right)^2\right] \times 10^2$$

$= \sigma^2 = 1.185 \times 100$

= 118.51

Mean = 25, Variance (σ^2) = 118.51.

Important

Coefficient of variation and variance are the different concepts.

Coefficient of variation is estimated as $\frac{\sigma}{\bar{x}} \times 100$; variance is simply the square of standard deviation.

(Variance = σ^2)

Coefficient of Variation

Coefficient of variation is 100 times the coefficient of dispersion based on standard deviation of a statistical series. It was first used by the famous statistician Karl Pearson. That is the reason why it is called Karl Pearson’s Coefficient of Variation. In the words of Karl Pearson, “Coefficient of variation is the percentage variation in the mean, the standard deviation being considered as the total variation in the mean. ”

In order to calculate coefficient of variation, standard deviation of the series is divided by mean of the series and multiplied by 100. It is estimated using the following formula:

FORMULA

Coefficient of Variation or CV

$$= \frac{\sigma}{\overline{X}} \times 100$$

= Coefficient of Standard Deviation × 100

Coefficient of Variation (CV) is 700 times the coefficient of dispersion based on standard deviation of a statistical series.

Coefficient of variation is used to compare the variability, homogeneity, stability and uniformity of two different statistical series. Higher value of coefficient of variation suggests greater degree of variability and lesser degree of stability. On the other hand, a lower value of coefficient of variation suggests lower degree of variability and higher, degree of stability, uniformity, homogeneity and consistency.

Calculation of Coefficient of Variation

(1) Individual Series and Coefficient of Variation

Illustration.

Calculate coefficient of variation of the following series:

S. No.	1	2	3	4	5	6	7	8	9	10
Marks	53	58	25	30	54	42	32	48	46	52

Solution:

S. No.	**Marks** **(X)**	**Deviation from Mean** **(x = X - X)** **X= 44**	**Square of Deviation** **$(X - X)^a$ or x^2**
1	53	9	81
2	58	14	196
3	25	- 19	361
4	30	- 14	196
5	54	10	100
6	42	- 2	4
7	32	- 12	144
8	48	4	16
9	46	2	4
10	52	8	64
N = 10	**$\sum X = 440$**		**$\sum x^2=\sum(X-x)^2= 1,166$**

$$\bar{X} = \frac{\Sigma X}{N} = \frac{440}{10} = 44$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{\sum(X - \bar{X})^2}{N}}$$

$$= \sqrt{\frac{1{,}166}{10}} = \sqrt{116.6} = 10.8$$

$$CV = \frac{\sigma}{\bar{X}} \times 100 = \frac{10.8}{44} \times 100$$

$$= 24.55$$

Coefficient of Variation = 24.55.

(2) Discrete Series or Frequency Array and Coefficient of Variation

Following illustration explains the calculation of CV for discrete frequency series.

Illustration.

Calculate coefficient of variation of the following data:

Items	10	12	14	16	18	20	22
Frequency	4	6	10	15	9	4	2

Solution:

Items	**Frequency** **(f)**	**Deviation from Assumed Average** **(dx = X - A)** **A=16**	**fdx**	**fdx^2**
10	4	-6	-24	144
12	6	-4	-24	96
14	10	-2	-20	40
16	15	0	0	0
18	9	2	+ 18	36
20	4	4	+ 16	64

22	2	6	+ 12	72
	N = 50		**∑fdx=-22**	**$\sum$fdx^2=45 2**

$$\overline{\mathrm{X}} = \mathrm{A} + \frac{\Sigma\mathrm{fdx}}{\mathrm{N}} = 16 + \frac{-22}{50} = 1556$$

$$\sigma = \sqrt{\frac{\sum\mathrm{fdx}^2}{\mathrm{N}} - (\frac{\sum\mathrm{fdx}}{\mathrm{N}})^2}$$

$$= \sqrt{\frac{452}{50} - (\frac{-22}{50})^2}$$

$$= \sqrt{904 - (-0)44)^2}$$

$$= \sqrt{904 - 01936}$$

$$= \sqrt{88464}$$

$$= 2.97$$

$$\mathrm{CV} = \frac{\sigma}{\overline{\mathrm{x}}} \times 100 = \frac{297}{1556} \times 100$$

$$= 19.09$$

Coefficient of Variation = 19.09.

(3) Frequency Distribution Series and Coefficient of Variation

illustration.

Calculate coefficient of variation, given the following data-set:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of Students	2	4	5	9	10	5	15

Solution:

Marks	**Mid-value** **(m)**	**Frequency** **(f>**	**Deviation from**	$dx' = \frac{dx}{C}$ **C = 10**	**dx'^{2}**	**fdx'**	**fdx'^{2}**

			Assumed Average **(dx = m - A) A = 35**				
0-10	5	2	-30	-3	9	- 6	18
10-20	15	4	- 20	-2	4	- 8	16
20-30	25	5	- 10	- 1	1	-5	5
30-40	35 (A)	9	0	0	0	0	0
40-50	45	10	+ 10	+ 1	1	+ 10	10
50-60	55	5	+ 20	+ 2	4	+ 10	20
60-70	65	15	+ 30	+ 3	9	+45	135
		N = 50				**∑fdx' = 46**	**∑fdx'²= 204**

$$\bar{X}A + \frac{\sum fdx'}{N} \times C$$

$$= 35 + \frac{46}{50} \times 10 = 44.2$$

$$\sigma = \sqrt{\frac{\sum fdx'^2}{N} - \left(\frac{\sum fdx'}{N}\right)^2} \times C$$

$$= \sqrt{\frac{204}{50} - (\frac{46}{50})^2} \times 10$$

$$= \sqrt{408 - (092)^2} \times 10$$

$$= \sqrt{408 - 08464} \times 10$$

$$= \sqrt{32336} \times 10$$

$$= 1.798 \times 10 = 17.98$$

$$\text{CV} = \frac{\sigma}{\overline{\text{X}}} \times 100$$

$$= \frac{17.98}{44.2} \times 100 = 40.68$$

Coefficient of Variation = 40.68.

Illustration.

Batsmen X and Y score following runs in different innings they played in a test series. Which of the two is a better scorer? Who is more consistent?

X	12	115	6	73	7	19	119	36	84	29
Y	47	12	76	42	4	51	37	48	13	0

Solution:

In order to determine which of the two is a better scorer, we should compare average runs scored by two batsmen in different innings. And, in order to determine consistency of batting we should compare CV of the runs scored by the batsmen in different series.

Batsman-X			**Batsman-Y**		
X	**(x=X - $\bar{X}$)** **$\bar{X}$ = 50**	**x^2**	**Y**	**(y=Y- $\bar{Y}$)** **$\bar{Y}$ = 33**	**y^2**
12	-38	1,444	47	+ 14	196
115	+ 65	4,225	12	-21	441
6	-44	1,936	76	+ 43	1,849
73	+ 23	529	42	+ 9	81
7	-43	1,849	4	-29	841
19	-31	961	51	+ 18	324
119	+ 69	4,761	37	+ 4	16
36	- 14	196	48	+ 15	225

84	+ 34	1,156	13	-20	400
29	-21	441	0	-33	1,089
∑X=500		**∑x²= 17,498**	**∑Y=330**		**∑y² = 5,462**

Batsman X: $\overline{\mathrm{X}} = \frac{\Sigma \mathrm{X}}{\mathrm{N}} = \frac{500}{10} = 50$

$$\sigma_{\mathrm{X}} = \sqrt{\frac{\Sigma \mathrm{x}^2}{\mathrm{N}}} = \sqrt{\frac{17{,}498}{10}} = \sqrt{1{,}749.8} = 41.83$$

and $\mathrm{CV_x} = \frac{\sigma}{\bar{\bar{X}}} \times 100 = \frac{41.83}{50} \times 100 = 83.1_{\grave{1}}6$

Batsman Y: $\overline{\mathrm{Y}} = \frac{\Sigma \mathrm{Y}}{\mathrm{N}} = \frac{330}{10} = 33$

$$\sigma_{\mathrm{Y}} = \sqrt{\frac{\Sigma \mathrm{y}^2}{\mathrm{N}}} = \sqrt{\frac{5{,}462}{10}} = \sqrt{546.2} = 23.37$$

$$\mathrm{CV_Y} = \frac{\sigma}{\overline{\overline{\mathrm{X}}}} \times 100\mathrm{X}$$

$$= \frac{23.37}{33} \times 100 = 70.82$$

Since average score of X is more than that of Y, we conclude that X is a better batsman. But batting of X shows greater CV than Y. Hence, Y is relatively more consistent in batting than X.

Illustration.

Two factories A and B are located in some Industrial estate. Average wage and its standard deviation are given below separately for A and B. Find out coefficient of variation.

Factory	Average Weekly Wage	S.D.	Number of Workers
A	35	5	476
B	30	10	524

Solution:

Coefficient of Variation of Factory A	Coefficient of Variation of Factory B

$\text{CV}_\text{A} = \frac{\sigma}{\overline{\text{X}}} \times 100$ $= \frac{5}{35} \times 100 = 14.29$	$\text{CV}_\text{B} = \frac{\sigma}{\overline{\text{X}}} \times 100$ $= \frac{10}{30} \times 100 = 33.33$

A Mathematical Property of Standard Deviation

The sum of the squares of the deviations of items from the arithmetic mean is minimum.

Deviations taken from Mean			**Deviations from Assumed Mean**		
X	**x = X- $\bar{X}$** **X = 6**	**x^2**	**X**	**d = X-A**	**d^2**
2	-4	16	2	-2	4
4	-2	4	4(A)	0	0
6	0	0	6	+ 2	4
8	+ 2	4	8	+ 4	16
10	+ 4	16	10	+ 6	36
∑X = 30		**$\sum x^2$ = 40**			$\Sigma d^2 = 60$

Merits and Demerits of Standard Deviation

Merits

(1) **Based on all Values:** The calculation of standard deviation is based on all the values of a series. It does not ignore any value. Accordingly, it is a comprehensive measure of dispersion.

(2) **A Certain Measure:** Standard deviation is a clear and certain measure of dispersion. Therefore, it can be used in all situations.

(3) **Little Effect of a change in Sample:** Change in sample causes little effect on standard deviation. This is because deviation is based on all the values of a sample.

(4) **Algebraic Treatment:** Standard deviation is capable of further algebraic treatment.

Demerits

(1) **Difficult**: It is difficult to calculate and make use of standard deviation as a measure of dispersion.

(2) **Importance to Extreme Values:** In the calculation of standard deviation, extreme values tend to get greater importance.

LORENZ CURVE

Lorenz curve is another important measure of variability of the statistical series. This curve was first used by Max Lorenz. Hence, it is called Lorenz Curve. These curves are generally used to measure variability in the distribution of income and wealth.

Lorenz curve is a measure of deviation of actual distribution from the line of equal distribution. This is a cumulative percentage curve. The extent of deviation of the actual distribution from the equal distribution is called Lorenz coefficient. Greater the distance of

Lorenz curve from the line of equal distribution more is the inequality or variability in its series. On the other hand, closer is the Lorenz curve to the line of equal distribution, lower will be the variability or degree of *inequality.

Construction of Lorenz Curve

Following steps are involved in the construction of a Lorenz Curve:

(i) First of all, the series is converted into a cumulative frequency series. The cumulative sum of the items (or mid-values of class intervals) is assumed to be 100 and the different items are converted into percentages of the cumulative sum. Likewise, cumulative sum of the frequencies is assumed to be 100 and different frequencies are converted into percentages of the sum of the frequencies.

(ii) Cumulative frequencies are plotted on X-axis of a graph, while cumulative items are plotted on the Y-axis.

(iii) On both axes, we start from 0 to 100 and both X and Y axes take the values from 0 to 100.

(iv) Draw a diagonal line joining the origin (0, 0) with the cumulative frequencies (100, 100). This shows equal distribution. It is, therefore, called 'Equality Line' or Line of Equal Distribution.

(v) Actual data are plotted on the graph and a curve is obtained by joining different points. This curve shows actual distribution.

(vi) The actual distribution curve is called Lorenz curve. Closeness of Lorenz curve to the Equal Distribution Line shows lesser variation in the distribution. Larger the gap between the actual distribution curve and the Lorenz Line, greater is the variation. If two Lorenz curves are drawn on the same graph paper, the one which is further away from the equal distribution line shows greater variation.

What is Lorenz Curve?

It is a curve that shows deviation of actual distribution (of income or wealth) from the line showing equal distribution.

Illustration.

Draw a Lorenz curve of the data given below:

Income (Rs.)	100	200	400	500	800
Number of Persons	80	70	50	30	20

Solution:

Income (X)	Cumulative Sum	Cumulative Percentage	Number of Persons (Frequency)	Cumulative Frequency	Cumulative Percentage of Frequency
100	100	5	80	80	32
200	300	15	70	150	60
400	700	35	50	200	80
500	1,200	60	30	230	92
800	2,000	100	20	250	100
∑X = 2,000			∑f = 250		

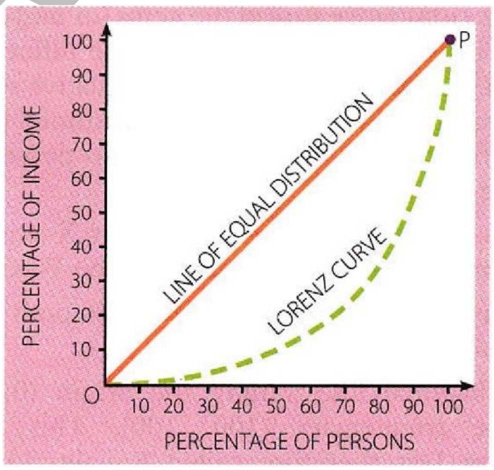


Illustration.

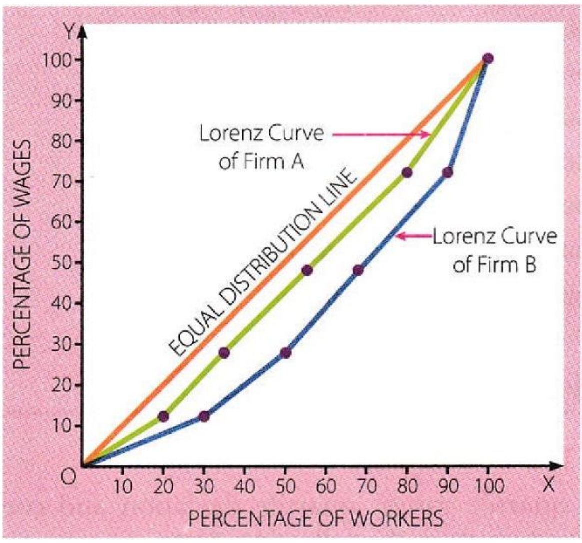
Show inequality in wages in two different firms using Lorenz Curve approach, given the following data:

Wages (Rs.)	50-70	7o-eo	90-110	110-130	130-150
Number of Workers A	20	15	20	25	20
Number of Workers B	150	100	90	110	50

Solution:

Estimation of Cumulative Sum and Percentage

Wages (Rs.)	Mid-value (X)	Cumulative Sum	Cumulative %	Firm A			Firm B		
				Number of Workers or Frequency	Cumulative Frequency	Cumulative %	Number of Workers or Frequency	Cumulative Frequency	Cumulative %
50-70	60	60	12	20	20	20	150	150	30
70-90	80	140	28	15	35	35	100	250	50
90-110	100	240	48	20	55	55	90	340	68
110-130	120	360	72	25	80	80	110	450	90
130-150	140	500	100	20	100	100	50	500	100
	∑X=500						1		



Observation: Both in firm A and firm B distribution of wages is far from equal. However, inequality is more pronounced in case of firm B than firm A as the Lorenz curve for firm B is farther away from the line of equal distribution.

Learning by doing

The following table shows number of firms in two different areas A' and 'B' according to their annual profits. Present the data by way of Lorenz curve.

Profit ('000' Rs.) 6	25	60	84	105	150	170	400
Firms in Area A 6	11	13	14	15	17	10	14
Firms in Area B 2	38	52	28	38	26	12	4

Make a Lorenz curve of the following data:

Income	500	1,000	2,000	3,000	3,500
Number in Class 'A' (000)	4	6	8	12	10
Number in Class 'B' (000)	8	7	5	3	2

Application of Lorenz Curve

Lorenz Curve is a graphic measure of dispersion in a statistical series. It is a very simple measure and provides an immediate glimpse of the degree of variation in a statistical distribution from its mean value. It was first used by **Prof. Lorenz** for the measurement of economic inequality relating to the distribution of income and wealth across different nations or across different periods of time for the same nation. With the passage of time the application of Lorenz curve has widely spread to measure disparity of distribution relating to various parameters like distribution of profits and wages. Briefly, Lorenz curve as a measure of dispersion is presently applied to the following parameters, viz.,

(i) distribution of income,

(ii) distribution of wealth,

(iii) distribution of wages,

(iv) distribution of profits,

(v) distribution of production, and

(vi) distribution of population.

Multiple Choice Questions

Select the correct alternative:

1. Which is the relative measure of dispersion?

(a) Range

(b) Mean deviation

(c) Coefficient of standard deviation

(d) None of these

2. Coefficient of range is:

(a) $\left(\frac{H+L}{H-L}\right) \times 2$

(b) $\frac{\mathrm{H+L}}{2}$

(c) $\frac{H+L}{H-L}$

(d) $\frac{H-L}{H+L}$

3. Which of the following formulae is used to find out inter quartile range?

(a) $\frac{\mathrm{a_1 - Q_3}}{2}$

(b) $\frac{\mathrm{Q_1 + Q_3}}{2}$

(c) $\mathrm{Q_3 - Q_1}$

(d) $\mathrm{Q_3 + Q_1}$

4. Quartile deviation is equal to:

(a) $\frac{Q_1 - Q_3}{2}$ (b) $\frac{Q_1 + Q_3}{2}$

(c) $\frac{Q_3 - O_1}{2}$ (d) $\frac{Q_3 + Q_1}{2}$

5. Mean deviation can be calculated by using:

(a) mean

(b)mode

(c) median

(d) all of these

6. Coefficient of mean deviation from mean is:

(a) $\frac{MD_{\bar{x}}}{\bar{X}}$

(b) $\frac{MD_m}{M}$

(c) $\frac{MD_2}{2}$

(d) all of these

7. Formula of standard deviation is:

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

(c) $\sigma = \sqrt{\frac{\sum(X - \bar{X})}{N}}$ (d) $\sigma = \sqrt{\frac{\sum X}{N}}$

8. Coefficient of variation is a percentage expression of:

(a) mean deviation

(b)quartile deviation

(c) standard deviation

(d)none of these

9. Which of these is the merit of standard deviation?

(a) Standard deviation is based on all values of the series

(b) Standard deviation shows little effect of changes in the sample

(c) In the estimation of standard deviation, more importance is given to difficult and extreme value

(d) Both (a) and (b)

10. $\sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$ is the formula of:

(a) combined mean deviation

(b) combined quartile deviation

(c) combined standard deviation

(d) coefficient of variation

11. In the calculation of standard deviation, deviations are taken only from the value of the series.

(a) mean

(b) mode

(c) median

(d) quartile

12. Which of the following equations is correct?

(a) Variance = σ

(b) Variance = σ^2

(c) Variance = σ^4

d) Variance = $\sqrt{\sigma}\times 2$

13. The standard deviation of a set of 50 observations is 8. If each observation is multiplied by 2, then the new value of standard deviation will be:

(a) 4

(b) 8

(c) 16

(d) none of the above

14. The standard deviation of a set of 50 observations is 6.5. If value of each observation is increased by 5, then the standard deviation is:

(a) 2.5

(b) 6.5

(c) 3.5

(d) none of the above

15. If the values of a set are measured in cm, the unit of variance will be:

(a) Cm

(b) No unit

(c) Cm2

(d) Cm3

16. A set of values is said to be relatively uniform if it has:

(a) High dispersion

(b) Zero dispersion

(c) Little dispersion

(d) Negative dispersion

17. Which one is difficult to compute?

(a) Relative measures of dispersion

(b) Absolute measures of dispersion

(c) Both (a) and (b)

(d) Range

18. If the variables are increased or decreased by the same proportion, the standard deviation changes by:

(a) Same proportion

(b) Different proportion

(c) Both (a) and (b)

(d) None of these

19. The most commonly used measure of dispersion is

(a) Coefficient of variation

(b) Standard deviation

(c) Range

(d) Quartile deviation

20. “Root-Mean Square Deviation from Mean” is

(a) Standard deviation

(b) Quartile deviation

(c) Both (a) and (b)

(d) None of these

21. Which of the following is not a measure of dispersion?

(a) Variance

(b) Mean deviation

(c) Standard deviation

(d) Mode

22. Which of the following is a relative measure of dispersion?

(a) Standard deviation

(b) Variance

(c) Coefficient of variation

(d) All of the above

23. Which of the following is a unitless measure of dispersion?

(a) Standard deviation

(b) Mean deviation

(c) Coefficient of variation

(d) Range

24. The standard deviation of 25 numbers is 40. If each of the numbers is increased by 5, then the new standard deviation will be:

(a) 40

(b) 45

(c) 41.5

(d) None of these

25. Semi-interquartile range is also known as

(a) Mean deviation

(b) Standard deviation

(c) Quartile deviation

(d) Quartile range.

26. For comparison of two different series, the best measure of dispersion is:

(a) Mean deviation

(b) Range

(c) Standard deviation

(d) Coefficient of variation

27. If the minimum value in a set is 9 and its range is 57, the maximum value of the set is:

(a) 33

(b) 66

(c) 48

(d) None of these

28. If mean and coefficient of variation of a set of data is 10 and 5, respectively, then the standard deviation is:

(a) 10

(b) 0.5

(c) 5

(d) none of the above

29. If all the observations are multiplied by 5, then

(a) New S.D. would be also multiplied by 5

(b) New S.D. would be increased by 5

(c) New S.D. would be half of the previous S.D

(d) New S.D. would be decreased by 5

30. The measure of dispersion which ignores signs of the deviations from a central value is:

(a) Quartile deviation

(b) Range

(c) Standard deviation

(d) Mean deviation

31. Which measure of dispersion has a different unit other than the unit of measurement of values?

(a) Mean deviation

(b) Range

(c) Standard deviation

(d) Variance

32. Which measure of dispersion is based on the absolute deviations only?

(a) Quartile deviation

(b) Mean Variation

(c) Standard deviation

(d) Range

33. Which measure of dispersion is not affected by the presence of extreme observations?

(a) Standard deviation

(b) Mean deviation

(c) Range

(d) Quartile deviation

34. The range of the following set of observations 2, 3, 5, 9, 8, 7, 6, 5, 7, 4, 3 is:

(a) 6

(b) 7

(c) 5.5

(d) 11

35. Which measures of dispersion is the quickest to compute?

(a) Mean deviation

(b) Quartile deviation

(c) Standard deviation

(d) Range

36. Which one is an absolute measure of dispersion?

(al Standard Deviation

(b) Mean Deviation

(c) Range

(d) All these measures

37. Which of the following measures of dispersion can attain a negative value?

(a) Range

(b) Mean deviation

(c) Standard deviation

(d) Variance

38. The measure of variation which is mostly affected by extreme items is:

(a) Range

(b) Quartile deviation

(c) Standard deviation

(d) Mean deviations

39. Coefficient of variation is

(a) Absolute measure

(b) Relative measure

(c) Both (a) and (b)

(d) None of these

40 The range represents the

(a) Difference between highest and lowest value

(b) Middle number

(c) Highest number

(d) Lowest number

41. The appropriate measure of dispersions for open - end classification is:

(a) Mean deviation

(b) Standard deviation

Quartile deviation

(d) All these measures

42. Quartile deviation is called

(a) Inter quartile range

(b) Quartile range

(c) Both (a) and (b)

(d) None of these

43. If the first quartile is 104 and quartile deviation is 8, the third quartile is:

(a) 130

(b) 120

(c) 136

(d) 146

44. Which measure is based on only the central fifty per cent of the observations?

(a) Mean deviation

(b) Quartile deviation

(c) Standard deviation

(d) All these measures

45. The square of standard deviation is known as:

(a) Variance

(b) Mean deviation

(c) Standard deviation

(d) None of these

46. When it comes to comparing two or more distributions, we consider

Relative measures of dispersion

(b) Absolute measures of dispersion

(c) Both (a) and (b)

(d) Either (a) or (b)

47. Standard Deviation is

(a) Absolute measure

(b) Relative measure

Both (a) and (b)

(d) None of these

Answers

Multiple Choice Questions

1(c)	**2. (d)**	**3. (c)**	**4. (c)**
5. (d)	**6. (a)**	**7. (b)**	**8. (c)**
9. (d)	**10. (c)**	**11. (a)**	**12. (b)**
13. (c)	**14. (b)**	**15. (c)**	**16. (c)**
17. (a)	**18. (d)**	**19. (b)**	**20. (a)**
21. (d)	**22. (c)**	**23. (c)**	**24. (a)**
25. (c)	**26. (d)**	**27. (b)**	**28. (b)**
29. (a)	**30. (d)**	**31. (d)**	**32. (b)**
33. (d)	**34. (b)**	**35. (d)**	**36. (d)**
37. (a)	**38. (a)**	**39. (b)**	**40. (a)**
41. (c)	**42. (a)**	**43. (b)**	**44. (b)**
45. (a)	**46. (a)**	**47. (a)**	

CHAPTER 12

CORRELATION

What is Correlation?

It is a statistical method or a statistical technique that measures quantitative relationship between different variables, like between price and demand.

1. CONCEPT AND DEFINITION OF CORRELATION

The statistical methods so far studied in this book focus on the analysis of one variable or one statistical series only. In real life however, two or more than two statistical series may be found to be mutually related. For instance, change in price leads to change in quantity demanded. Increase in supply of money causes increase in price level. Increase in level of employment results in increase in output. Such situations necessitate simultaneous study of two or more statistical series. The focus of study in such situations is on the degree of relationship between different statistical series. The statistical technique that studies the degree of such relationships is called the technique of correlation.

Definition

According to Croxton and Cowden, "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

In the words of Boddington, "Whenever some definite connection exists between the two or more groups, classes or series or data there is said to be correlation. "

Relationship between Two Variables may just be a Coincidence

One may find a relationship between two variables which is just a coincidence. Example: When there is a departure of migratory birds from a sanctuary, you may find a fall in wedding ceremonies in the country. Such relationships are meaningless. These are in other words, spurious relationships which are devoid of any meaningful conclusion. Such relationships are not to be treated as correlations, Only those relationships are to be treated as correlations which offer some meaningful conclusions.

Example: Increase in rainfall and increase in rice production is a relationship that makes sense: increase in per capita income and decrease in death rate is a meaningful relationship; Good percentage of marks in physics may be related to good percentage of marks in mathematics; and so on.

Positive and Negative Correlation

Correlation between different variables may either be positive or negative. Here is a brief description of the two:

(1) Positive Correlation

When two variables move in the same direction, that is, when one increases the other also increases and when one decreases the other also decreases, such a relation is called positive correlation.

Relationship between price and supply may be cited as an example. Check the following table as an illustration:

Table 1. Positive Correlation between the Variables

(A) Simultaneous Increase in the Values of both Variables		**(B) Simultaneous Decrease in the Values of both Variables**	
X	Y	X	Y
10	100	50	200
20	150	40	150
30	200	30	100
40	250	20	50

(2) Negative Correlation

When two variables change in different directions, it is called negative correlation. Relationship between price and demand, may be cited as an example. The following table demonstrates this relation:

Table 2. Negative Correlation between the Variables

Rise in the Value of One Variable is	**accompanied with a fall in the other**
X	**Y**
1	5
2	4
4	2

5	1

What does Correlation Measure?

Often the students tend to believe that correlation suggests a relationship between two variables where one is the cause of the other. Example: This is correlation between price and quantity demanded of a commodity. Clearly, an increase in price causes a decrease in quantity demanded, and vice versa. Change in price causes changes in quantity demanded. But to be emphatically noted is the point that cause and effect relationship between the variables is not at all any pre-condition in the theory of correlation. Correlation just measures the degree and intensity of relationship between the two variables, with or without any cause and effect relationship. Of course, the established relationship between the variables should be capable of offering us some meaningful conclusion. Example: Students who are good in academics may be good in sports also. Certainly, it is a meaningful relationship (or correlation) if one finds it. But surely there is no cause and effect relationship between the two variables.

Linear and Non-Linear Correlation

Correlation may be linear or non-linear.

(1) Linear Correlation

When two variables change in a constant proportion, it is called linear correlation. If the two sets of data bearing fixed proportion to each other are shown on a graph paper, their relationship will be indicated by a straight line. Thus, linear correlation implies a straight-line relationship.

(2) Non-linear Correlation

When the two variables do not change in any constant proportion, the relationship is said to be non-linear. Such a relationship does not form a straight-line relationship.

Illustration.

Linear Correlation

(a)	2	4	6	8	10	12	14
(b)	5	10	15	20	25	30	35

Thus, for every change in variable (a) by 2 units there is a change in variable (b) by 5 units.

Non-linear Correlation

(a)	2	4	6	8	10	12	14

(b)	3	7	12	18	25	35	45

Here there is no specific relationship between the two variables, though both tend to change in the same direction. That is, both are increasing, but not in any constant proportion.

Simple and Multiple Correlation

(1) Simple Correlation

Simple correlation implies the study of relationship between two variables only. Like the relationship between price and demand or the relationship between money supply and price level.

(2) Multiple Correlation

When the relationship among three or more than three variables is studied simultaneously, it is called multiple correlations. In case of such correlation, the entire set of independent and dependent variables is simultaneously studied. For instance, effects of rainfall, manure, water, etc., on per hectare productivity of wheat are simultaneously studied.

The basic difference between Linear and Non-linear Correlation

In case of linear correlation, the two sets of data show some fixed proportion to each other and therefore, form a straight line on a graph paper, in case of nonlinear correlation, the two sets of data do not show any fixed proportion to each other, and therefore, do not make a straight line on a graph paper.

Partial Correlation

When more than two variables are involved and out of these the relationship between only two variables is studied treating other variables as constant, then the correlation is partial.

Correlation and Causation

Correlation is a numerical measure of direction and magnitude of the mutual relationship between the values of two or more variables. But the presence of correlation should not be taken as the belief that the two correlated variables necessarily have causal relationship as well. Correlation does not airways arise from causal relationship but with the presence of causa relationship, correlation is certain to exist.

2. DEGREE OF CORRELATION

Degree of correlation refers to the Coefficient of Correlation. There can be the following degrees of positive and negative correlation.

(1) Perfect Correlation: When two variables change in the same proportion it is called perfect correlation. It may be of two kinds:

(i) Perfect Positive: Correlation is perfectly positive when proportional change in two variables is in the same direction. In this case, coefficient of correlation is positive (+1).

(ii) Perfect Negative: Correlation is perfectly negative when proportional change in two variables is in the opposite direction. In this case, coefficient of correlation is negative (-1).

(2) **Absence of Correlation:** If there is no relation between two series or variables, that is, change in one has no effect on the change in other, then those series or variables lack any correlation between them.

(3) **Limited Degree of Correlation:** Between perfect correlation and absence of correlation there is a situation of limited degree of correlation. In real life, one mostly finds limited degree of correlation. Its coefficient (r) is more than zero and less than one (r > 0 but < 1). The degree of correlation between 0 and 1 may be rated as:

(i) **High:** When correlation of two series is close to one, it is called high degree of correlation. Its coefficient lies between 0.75 and 1,

(ii) Moderate: When correlation of two series is neither large nor small, it is called moderate degree of correlation. Its coefficient lies between 0.25 and 0.75.

(iii) Low: When the degree of correlation of two series is very small, it is called low degree of correlation. Its coefficient lies between 0 and 0.25.

All these degrees of correlations may be positive or negative.

Degree of Correlation

Degree	**Positive**	**Negative**
Perfect	+ 1	-1
High	Between + 0.75 and +1	Between - 0.75 and -1
Moderate	Between + 0.25 and + 0.75	Between - 0.25 and - 0.75
Low	Between 0 and + 0.25	Between 0 and - 0.25
Zero	o	0

3. METHODS OF ESTIMATING CORRELATION

Various methods are available for estimating correlation between different sets of statistical series. Some of the important ones are as under:

(1) Scattered Diagram Method,

(2) Karl Pearson's Coefficient of Correlation, and

(3) Spearman's Rank Correlation Coefficient.

Line of Best Fit

Line of Best Fit is the one that passes through the scattered points such that it represents most of these points. Roughly, half of the scattered points should be on either side of the line.

Scattered Diagram

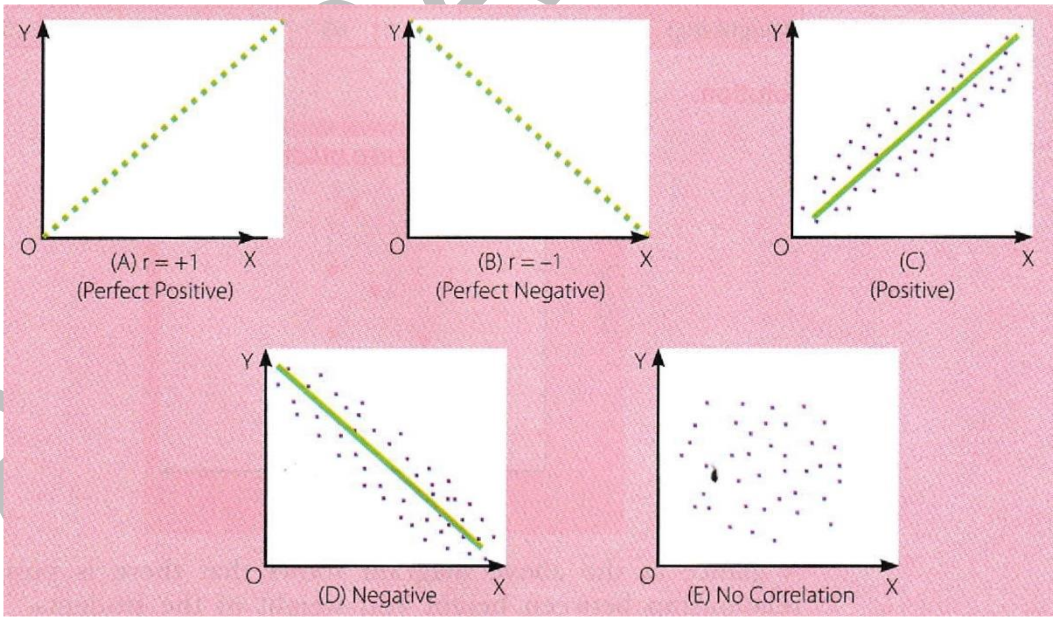
Scattered diagram offers a graphic expression of the directum and degree of correlation. To make a Scattered Diagram, data are plotted on a graph paper. A dot is marked for each value. The course of these dots would indicate direction and closeness of the variables. Following pictures show some of the possible directions and the degrees of closeness of the variables.

As should be clear from the scattered diagrams, closeness of the dots towards each other in a particular direction indicates higher degree of correlation. If the dots are scattered (showing neither the closeness nor any direction), it is an indication of low degree of correlation.

A Note

Scattered diagram only shows an approximation of the relationship or closeness of two sets of data. Precise measurement of the relationship is not possible.

Graph: Scattered Diagram



Merits and Demerits of Scattered Diagram Merits

(i) Scattered diagram is a very simple method of studying correlation between two variables.

(ii) Just a glance at the diagram is enough to know if the values of the variables have any relation or not.

(iii) Scattered diagram also indicates whether the relation is positive or negative.

Demerits

(i) A scattered diagram does not measure the precise extent of correlation.

(ii) It gives only an approximate idea of the relationship.

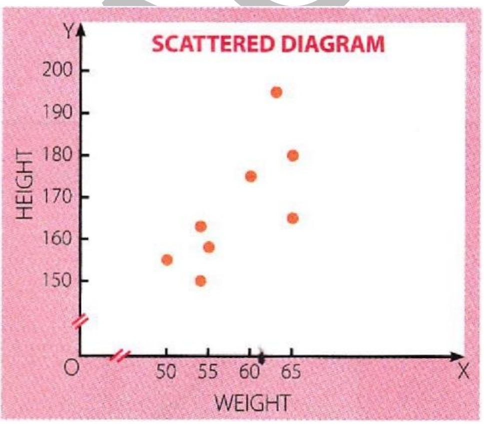
(iii) It is not a quantitative measure of the relationship between the variables. It is only a qualitative expression of the quantitative change.

Illustration.

The following table gives height and weight of the students of a class. Make a scattered diagram to show if the relationship is positive or negative and if the relationship is strong or weak.

Height (cm)	180	150	158	165	175	163	195	155
Weight (kg)	65	54	55	65	60	54	63	50

Solution:



A glance at the above diagram shows that there is positive relationship between height and weight of the students. The dots are moving upward in a particular course from left to right. It shows that with the increase in height, weight also increases.

However, this is a case of limited positive correlation, because the dots do not make any straight line.

Karl Pearson's Coefficient of Correlation

Scattered diagram method of correlation merely indicates the direction of correlation but not its precise magnitude. Karl Pearson has given a quantitative method of calculating correlation. It is an important and widely used method of studying correlation. Karl Pearsons’ coefficient of correlation is generally written as V.

FORMULA

According to Karl Pearson’s method, the coefficient of correlation is measured as:

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$$

where,

r = Coefficient of correlation,

x = X- $\bar{X}$.

y =Y- $\bar{Y}$.

σ_x = Standard deviation of X series,

σ_y = Standard deviation of Y series.

N = Number of observations.

This formula is applied only to those series where deviations are worked out from actual average of the series, it does not apply to those series where deviations are calculated on the basis of assumed mean. Value of the coefficient of correlation calculated on the basis of this formula may vary between +1 and -1.

However, the situations, when r = +1, r =-l, or r = 0, are rather rare. Generally, value of V varies between + 1 and - 1.

Note

Karl Pearson's coefficient of correlation does not apply to those series where deviations are calculated on the basis of assumed mean.

A Modified Version of Karl Pearson's Formula

In it there is no need to calculate standard deviation of ‘X’ and ‘Y\ Coefficient of correlation may be worked out directly using the following formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

Here, x = (X - X̄); y = (Y - ȳ).

Illustration.

Calculate coefficient of correlation, given the following data:

X	2	3	4	5	6	7	8
Y	4	7	8	9	10	14	18

Solution:

Calculation of Coefficient of Correlation

X	Deviation (x = X- $\bar{X}$) $\bar{X}$ = 5	Square of Deviation (x^2)	Y	Deviation (y = Y-$\bar{Y}$) $\bar{Y}$ = 5	Square of Deviation (y^2)	Multiple of Deviations (xy)
2	-3	9	4	-6	36	+ 18
3	- 2	4	7	-3	9	+ 6
4	- 1	1	8	-2	4	+ 2
5	0	0	9	- 1	1	0
6	1	1	10	0	0	0
7	2	4	14	4	16	+ 8
8	3	9	18	8	64	+24
∑X = 35	∑x = 0	$\sum x^2$ = 28	∑Y = 70	∑y = 0	$\sum y^2$ = 130	∑xy = 58
N = 7			N = 7			
X = 5			Y = 10			

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}; x = (X - \bar{X}); y = (Y - \bar{Y})$$

The table show's that, $\Sigma xy = 58, \Sigma x^2 = p8, \Sigma y^2 = 130$.

Substituting the values, we get

$$r = \frac{58}{\sqrt{28 \times 130}} = \frac{58}{\sqrt{3,640}} = \frac{58}{60.33} = +0.96$$

Coefficient of Correlation (r) = + 0.96.

It is a situation of high positive correlation.

Illustration.

Calculate the coefficient of correlation between the age of husbands and wives.

Age of Husband (Years)	21	22	28	32	35	36
Age of Wife (Years)	18	20	25	30	31	32

Solution:

Age of Husband (X)	**Deviation (x= X- $\bar{X}$) $\bar{X}$ = 29**	**Square of Deviation (x^2)**	**Age of Wife (Y)**	**Deviation (y= Y-$\bar{Y}$) $\bar{Y}$ = 26**	**Square of Deviation y^2**	**Multiple of Deviations (xy)**
21	-8	64	18	-8	64	64
22	-7	49	20	-6	36	42
28	- 1	1	25	- 1	1	1
32	+ 3	9	30	+ 4	16	12
35	+ 6	36	31	+ 5	25	30
36	+ 7	49	32	+ 6	36	42
∑X = 174	∑x = 0	$\sum x^2 = 208$	∑Y = 156	∑y = 0	$\Sigma y^2 = 178$	∑xy = 191

$x = (X - \bar{X}); y = (Y - \bar{Y})$

$X = \frac{\Sigma X}{N} = \frac{174}{6} = 29; Y^{-} = \frac{\Sigma Y}{N} = \frac{156}{6} = 26$

$\Sigma xy = 191, \Sigma x^2 = 208, \Sigma y^2 = 178$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$r = \frac{191}{\sqrt{208 \times 178}}$$

$$= \frac{191}{\sqrt{37{,}024}}$$

$$= \frac{191}{192.42} = 0.993$$

Coefficient of Correlation (r) = 0.993.

Thus, there is a high degree of positive correlation between the age of husband and wife.

Short-cut Method

This method is used when mean value is not in whole number but in fractions. In this method, deviation is calculated by taking the assumed mean of both the series. It involves the following steps:

(i) Any convenient value in X and Y series is taken as assumed mean A_x and A_y.

(ii) With the help of assumed mean of both the series, deviation of the values of individual variable, i.e., dx $(X - A_x)$ and dy $(Y - A_y)$ are calculated.

(iii) ∑dx and ∑dy are found by adding the deviations.

(iv) Deviations of the two series are multiplied, as dx.dy, and the multiples added up to obtain ∑dxdy.

(v) Squares of the deviations dx' and dy^2 are added up to find out $\sum dx^2$ and $\sum dy^2$.

(vi) Finally, coefficient of correlation is calculated using the following formula:

FORMULA

$$r = \frac{\Sigma dxdy - \frac{(\Sigma dx)\times(\Sigma dy)}{N}}{\sqrt{\sum dx^2 - \frac{(\Sigma dx)^2}{N}} \times \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

Here,

dx = Deviation of X series from the assumed mean = (X-A),

dy = Deviation of Y series from the assumed mean = (Y - A),

∑dxdy = Sum of the multiple of dx and dy.

$\sum dx^2$ = Sum of square of dx.

$\sum dy^2$ = Sum of square of dy.

∑dx = Sum of deviation of X series.

∑dy = Sum of deviation of Y series.

N = Total number of items.

Illustration. *

Calculate coefficient of correlation between the price and quantity supplied.

Price (Rs.)	4	6	8	15	20
Supply (kg)	10	15	20	25	30

Solution:

Price **(X)**	**Deviation** **(dx = X - A)** **A = 8**	**Square of Deviation (dx²)**	**Supply** **00**	**Deviation** **(dy = V - A)** **A = 20**	**Square of Deviation (dy²)**	**Multiple of Deviations (dxdy)**
4	-4	16	10	- 10	100	40
6	-2	4	15	- 5	25	10
8(A)	0	0	20(A)	0	0	0
15	7	49	25	5	25	35
20	12	144	30	10	100	120
N = 5	∑dx = 13	∑dx² = 213	N = 5	∑dy = 0	∑dy² = 250	∑dxdy = 205

$$r = \frac{\Sigma dxdy - \frac{(\Sigma dx)\times(\Sigma dy)}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}} \times \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

$$= \frac{205 - \frac{(13)\times(0)}{5}}{\sqrt{213 - \frac{(13)^2}{5}} \times \sqrt{250 - \frac{(0)^2}{5}}}$$

$$= \frac{205 - 0}{\sqrt{213 - 33.80} \times \sqrt{250 - 0}}$$

$$= \frac{205}{\sqrt{179.20} \times \sqrt{250}} = \frac{205}{13.39 \times 15.81}$$

$$= \frac{205}{211.70}$$

$$= +0.97$$

Coefficient of Correlation (r) = + 0.97

This is a situation of a high degree of positive correlation.

Step-deviation Method

The method involves the following steps:

(i) Repeat Step-1 and Step-2 of the short-cut method.

(ii) Now divide ‘dx’ and ‘dy’ by some common factor as $dx' = \frac{dx}{C_1}, dy' = \frac{dy}{C_2}$ = here C_2 is common factor for scries X and C_2 is common factor for series Y. And dx' and dy' are step-deviations.

(iii) ∑dx' and ∑dy' are found by adding the deviations.

(iv) Deviations of the two series are multiplied, as dx' × dy', and the multiples added up to obtain ∑dx'dy'.

(v) Squares of the deviations dx'^2 and dy'^2 are added up to find out $\sum dx'^2$ and $\sum dy'^2$.

(vi) Finally, coefficient of correlation is calculated using the following formula:

FORMULA

$$\frac{\sum dx'dy' - \frac{(\sum dx') \times (\sum dy')}{N}}{\sqrt{\sum dx'^2 - \frac{(\sum dx')^2}{N}} \times \sqrt{\sum dy'^2 - \frac{(\sum dy')^2}{N}}}$$

Illustration.

Calculate coefficient of correlation between the price and quantity demanded.

Price (Rs.)	5	10	15	20	25
Demand (kg)	40	35	30	25	20

Solution;

Price **(X)**	**dx =X- A_x** **A_x =15**	**$dx' = \frac{dx}{C_1}$** **C_1 = 5**	**dx'^2**	**Demand** **(Y)**	**dy = Y - Ay** **Ay = 30**	**$dy' = \frac{dy}{C_2}$** **C_2 = 5**	**dy'^2**	**dx'dy'**
5	- 10	-2	4	40	10	2	4	-4

10	- 5	- 1	1	35	5	1	1	- 1
15	0	0	0	30	0	0	0	0
20	5	1	1	25	-5	-1	1	- 1
25	10	2	4	20	-10	-2	4	-4
N = 5		∑dx' = 0	$\Sigma dx'^2 = 10$	N = 5		Σdy' $= 0$	∑dy'^{2} = 10	∑dx'dy'=-10

$$\frac{\Sigma dx'dy' - \frac{(\Sigma dx') \times (\Sigma dy')}{N}}{\sqrt{\Sigma dx'^2 - \frac{(\Sigma dx')^2}{N}} \times \sqrt{\Sigma dy'^2 - \frac{(\Sigma dy')^2}{N}}}$$

$$= \frac{-10 - \frac{0}{5}}{\sqrt{10 - \frac{0}{5}} \times \sqrt{10 - \frac{0}{5}}}$$

$$= \frac{-10}{\sqrt{10} \times \sqrt{10}} = \frac{-10}{10} = -1$$

Coefficient of Correlation (r) = - 1.

This is a situation of a perfectly negative correlation between price and quantity demanded.

Properties of Correlation Coefficient

(i) r has no unit. It is a pure number. It means units of measurement are not parts of r.

(ii) A negative value of r indicates an inverse relation, and if r is positive, the two variables move in the same direction.

(iii) If r = 0, the two variables are uncorrelated. There is no linear relation between them. However, other types of relation may be there.

(iv) If r = 1 or r = - 1, the correlation is perfect or proportionate. A high value of r indicates strong linear relationship, i.e., + 1 or -1.

(v) The value of the correlation coefficient lies between minus one and plus one, i.e., -l ≤ r ≤ + l .If the value of r lies outside this range, it indicates error in calculation.

Spearman's Rank Correlation Coefficient

In 1904, **Charles Edward Spearman** developed a formula to calculate coefficient of correlation of qualitative variables. It is popularly known as Spearman’s Rank Difference Formula or Method. There are some variables whose quantitative measurement is not possible. These variables are known as qualitative variables (or more precisely ‘attributes’) such as beauty, bravery, wisdom, ability, virtue, etc.

The attributes cannot be expressed in numbers or quantitative terms. Their relative merit can be determined on the basis of their order of preference or ranking. For instance, in a fancy-dress competition two judges may rank the participants in order of preference. Similarly, the selection committee may prepare a list of successful candidates in order of preference. We experience many such examples in our day-to-day life.

It is in such situations that we use Spearman’s Rank Difference Method.

FORMULA

$$r_k = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Here.

r = Coefficient of rank correlation.

D = Rank differences.

N = Number of pairs.

We shall study the calculation of rank correlation in three different situations:

(i) When Ranks are given;

(ii) When Ranks are not given; and

(iii) When the values of the series are the same.

Following illustration explains the calculation of Rank Correlation.

Coefficient of Rank Correlation when Ranks are given

Illustration.

Solution:

Judge I, X = R_1	**Judge II, Y= R_2**	$D = R_1 - R_2$	**D²**
1	10	-9	81
2	6	-4	16
3	5	-2	4
4	4	0	0
5	7	- 2	4

6	9	- 3	9
7	8	- 1	1
8	2	6	36
9	1	8	64
10	3	7	49
N = 10			ΣD^2 = 264

$$r_k = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Here, N $= 10; \Sigma \text{D}^2 = 264$

$$r_k = 1 - \frac{6 \times 264}{(10)^3 - 10}$$

$$= 1 - \frac{1{,}584}{990}$$

= 1 - 1.6 = - 0.6

Coefficient of Rank Correlation (r_k) = - 0.6.

There is a negative coefficient of rank correlation to the tune of -0.6 which is fairly high. High negative correlation suggests that two sets of judgements are fairly opposite to each other.

Coefficient of Rank Correlation when Ranks are not given

Sometimes, only data are given and not the ranks. In such situations, ranks are to be accorded by the student himself. While according the ranks, uniform procedure should be adopted for both the series. For example, if the highest rank is accorded to the lowest value in one series, the same must be done for the second series as well.

Illustration.

In a Poetry Recitation Competition, 10 participants were accorded following marks by two different judges, X and Y:

X	15	IV	14	13	11	12	16	18	10	9
Y	15	12	4	6	7	9	3	10	2	5

Calculate the coefficient of rank correlation.

Solution:

There are 10 items each in the two series in this question. We may accord, the highest rank, i.e., 10 to highest value/item in each series, i.e., 18 in series X and 15 in series Y, and accord the lowest rank, i.e., 1 to the lowest value/item in each series, i.e., 9 in series X and 2 in series Y.

X	Rank R_1	Y	Rank R_2	$D = R_1 - R_2$	D^2
15	7	15	10	-3	9
17	9	12	9	0	0
14	6	4	3	3	9
13	5	6	5	0	0
11	3	7	6	-3	9
12	4	9	7	- 3	9
16	8	3	2	6	36
18	10	10	8	2	4
10	2	2	1	1	1
9	1	5	4	~ 3	9
N = 10					$\Sigma D^2 = 86$

$$r_k = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Here, $N = 10; \Sigma D^2 = 86$

$$r_k = 1 - \frac{6 \times 86}{(10)^3 - 10}$$

$$= 1 - \frac{516}{1{,}000 - 10}$$

$$= 1 - 0.52$$

$$= 0.48$$

Coefficient of Rank Correlation (r_k) = 0.48.

Thus, there is a positive rank correlation of a moderate degree of 0.48.

Coefficient of Rank Correlation when Ranks are Equal

Sometimes, two or more items in the series have equal ranks. In such situations, average of the two ranks (say 7.5 of the ranks 7 and 8) is accorded to each value. But one is likely to commit mistake in this procedure. In order to avoid the possibility of error, the following formula is used for the calculation of coefficient of rank correlation in such situations:

FORMULA

$$r_k = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\right]}{N^3 - N}$$

Here, m = Number of items of equal ranks.

Illustration.

Calculate coefficient of rank correlation between the marks in Economics and Statistics, as indicated by 8 answer books of each of the two examiners.

Marks in Statistics	15	10	20	28	12	10	16	18
Marks in Economics	16	14	10	12	11	15	18	12

Solution:

There are 8 answer books each in Economics and Statistics indicating different marks. Rank 1 is accorded to the highest score. In Statistics, two answer books indicate 10 marks each. Hence, the first answer book has been given Rank 8 and the second 7. Thus, the average rank = $\frac{8+7}{2}$ = 7.5 has been accorded to both. Likewise, in Economics two answer books indicate 12 marks each. The average rank = $\frac{6+5}{2} = 5.5$ has, therefore, been accorded to both.

Calculation of Coefficient of Rank Correlation

Marks in Statistics (X)	Rank R_1	Marks in Economics (Y)	Rank R_2	D = R_1 - R_2	D^2
15	5	16	2	3.0	9.00
10	7.5	14	4	3.5	12.25
20	2	10	8	-6	36.00

28	1	12	5.5	- 4.5	20.25
12	6	11	7	- 1	1.00
10	7.5	15	3	4.5	20.25
16	4	18	1	3.0	9.00
18	3	12	5,5	- 2.5	6.25
N = 8					$\sum D^2$=114

Here, number 10 is repeated twice in series X and number 12 is repeated twice in series Y. Therefore, in X, m = 2 and in Y, m = 2.

$$r_k = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)]}{N^3 - N}$$

$$= 1 - \frac{6\left[114 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]}{8^3 - 8}$$

$$= 1 - \frac{6|114 + \frac{1}{12}(6) + \frac{1}{12}(6)}{512 - 8}$$

$$= 1 - \frac{6\left[114 + \frac{1}{2} + \frac{1}{2}\right]}{504}$$

$$= 1 - \frac{6[115]}{504} = 1 - \frac{690}{504}$$

$$= 1 - 1.36 = -0.36$$

Coefficient of Rank Correlation (r_k) = - 0.36.

Merits and Demerits of Rank Correlation

Merits

(i) Spearman’s rank correlation method is easier than Pearson’s method of correlation.

(ii) Rank correlation method is very convenient method when the series give only order of preference and not the actual values of the variables.

(iii) Rank correlation is a superior method of analysis in case of qualitative distributions such as those relating to virtue, wisdom or ignorance.

Demerits

(i) Rank correlation method cannot be used in case of group frequency distributions.

(ii) It can handle only a limited number of observations. It is generally not used when the number of observations exceeds 20.

4. IMPORTANCE OR SIGNIFICANCE OF CORRELATION

Following observations highlight the importance or significance of correlation as a statistical method:

(1) Formation of Laws and Concepts: The study of correlation shows the direction and degree of relationship between the variables. This has helped the formation of various laws and concepts in economic theory, such as, the law of demand and the concept of elasticity of demand.

(2) Cause and Effect Relationship: Correlation coefficient sometimes suggests cause and effect relationship between different variables. This helps in understanding why certain variables behave the way they behave.

(3) **Business Decisions:** Correlation analysis facilitates business decisions because the trend path of one variable may suggest the expected changes in the other. Accordingly, the businessman may plan his business decisions for the future.

(4) **Policy Formulation:** Correlation analysis also helps policy formulation. If the Government finds a negative correlation between tax rate and tax collection, it should pursue the policy of low tax rate. Because, low tax rate would lead to high tax collection.

Multiple Choice Questions

Select the correct alternative:

1. When two variables change in the same direction, then such a correlation is called:

(a) negative

(b) positive

(c) no correlation

(d) all of these

2. When the relation of three or more variables is studied simultaneously, it is called:

(a) simple correlation

(b) partial correlation

(c) multiple correlation

(d) none of these

3. Relation between price and demand is:

(a) positive

(b) negative

(c) one to one

(d) no relationship

4. When coefficient of correlation lies between +0.25 and + 0.75, it is called:

(a) perfect degree of correlation

(b) high degree of correlation

(c) moderate degree of correlation

(d) low degree of correlation

5. Coefficient of correlation lies always between:

(a) 0 and +1

(b) -1 and 0

(c)-1 and+1

(d) none of these

6. Rank correlation is a superior method of analysis in case of ____ distribution.

(a) qualitative

(b) quantitative

(c) frequency

(d) none of these

7. Which of the following equations is correct?

(a) $r_k = 1 - \frac{6\Sigma D^2}{N}$

(b) $r_k = 1 - \frac{6\Sigma D^2}{N^2 - N}$

(c) $r_k = 1 - \frac{6\Sigma D^2}{N^3 - N}$

(d) $r_k = 1 - \frac{6\Sigma D^2}{N^4 - N}$

8. Formula of Karl Pearson ' s coefficient of correlation is:

(a) $\frac{N\sigma_X\sigma_Y}{\sum xy}$

(b) $\frac{\sum xy}{N\sigma_x\sigma_Y}$

(c) $\frac{\sigma_X\sigma_Y}{\sum xy}$

(d) $\frac{\Sigma xy}{\sigma_X\sigma_Y}$

9. When two variables change in a constant proportion, it is called:

(a) linear correlation

(b) non-linear correlation

(c) partial correlation

(d) none of these

1. A scatter diagram:

(a; Is a statistical test

(b) Must be linear

(c) Must be curvilinear

(d) Is a graph of x and y values

2. Maximum value of rank correlation coefficient is:

(a) 0

(b) +1

(c) -1

(d) None of these

3. If the relationship between x and y is positive, as variable y decreases, variable x:

(a) Increases

(b) Decreases

(c) Remains same

(d) Changes linearly

4. The correlation coefficient will be -1 if the slope of the straight line in a scatter diagram is:

(a) Positive

(b) Negative

(c) Zero

(d) None of these

5. In a ‘negative’ relationship

(a) As x increases, y increases

(b) As x decreases, y decreases

(c) As x increases, y decreases

(d) Both (a) and (b)

6. If there is a perfect disagreement between the marks in Geography and Statistics, then what would be the value of rank correlation coefficient?

8. Scatter diagram helps us to:

(a) Find the nature of correlation between two variables

(b) Obtain the mathematical relationship between two variables

(c) Compute the extent of correlation between two variables

(d) Both (a) and (c)

9. The lowest strength of association is reflected by which of the following correlation coefficients?

(a) 0.95

(b) - 0.60

(c) -0.35

(d) 0.29

10. Kari Pearson’s coefficient is defined from:

(a) Ungrouped data

(b) Grouped data

(c) Both (a) and (b)

(d) None of these

11. When r = 1, all the points in a scatter diagram would lie:

(a) On a straight line directed from lower left to upper right

(b) On a straight line

(c) On a straight line directed from upper left to lower right

(d) Both (a) and (b)

12. The highest strength of association is reflected by which of the following correlation coefficients?

(a) -1.0

(b) -0.95

(c) 0.1

(d) 0.85

13. There is a high degree of negative correlation between ‘overweight’ and ‘life expectancy’. A correlation coefficient consistent with the above statement is:

(a) r = 0.80

(b)r = 0.20

(c) r = -0.20

(d) r = -0.80

14. Correlation coefficient is __________ of the units of measurement.

(a) Independent

(b) Dependent

(c) Both (a) and (b)

(d) None of these

15. The correlation between sale of cold drinks and day temperature is:

(a) Positive

(b) Negative

(c) Zero

(d) None of these

16. There is a high direct association between measures of ‘cigarette smoking’ and ‘lung damage’. The correlation coefficient consistent with the above statement is:

(a) 1

(b) Any value

(c) - 1

(d) (b) or (c)

7. The correlation between ages of husbands and wives is:

(a) Positive

(b) Negative

(c) Zero

(d) None of these

17. Simple correlation is called:

(a) linear correlation

(b) Nonlinear correlation

(c) Both (a) and (b)

(d) None of these

18. For finding the degree of agreement about beauty between two judges in a beauty contest, we use:

(a) Coefficient of correlation

(b) Coefficient of rank correlation

(c) Scatter diagram

(d) Coefficient of concurrent deviation

19. If all the plotted points in a scatter diagram lie on a single line, then the correlation is:

(a) Perfect positive

(b) Perfect negative

(c) Both (a) and (b)

(d) Either (a) or (b)

20. Correlation coefficient is dependent on the choice of both origin and the scale of observations,

(a) True

(b) False

(c) Both (a) and (b)

(d) none of these

21. The correlation between sale of woolen garments and day temperature is:

(a) Positive

(b) Negative

(c) Zero

(d) None of these

22. The correlation between shoe-size and intelligence is:

(a) Zero

(b) Negative

(c) Positive

(d) None of these

23. What are the limits of the correlation coefficient?

(a) -1 and 1

(b) No limit

(c) 0 and 1, including the limits

(d) -1 and 1, including the limits

24. If r is the correlation coefficients between the two variables, then:

(a) -1 ≤ r ≤ 0

(b) 1≤ r ≤ 2

(c) -1 ≤ r ≤ 1

(d) 0 ≤ r ≤ 1

Answers

Multiple Choice Questions

1.(b)	**2. (c)**	**3. (b)**	**4. (c)**
5. (c)	**6. (a)**	**7. (c)**	**8. (b)**
9. (a)	**10. (d)**	**11. (b)**	**12. (b)**
13. (b)	**14. (c)**	**15. (c)**	**16. (a)**
17. (a)	**18. (d)**	**19. (a)**	**20. (a)**
21. (a)	**22. (d)**	**23. (a)**	**24. (a)**
25. (b)	**26. (a)**	**27. (b)**	**28. (d)**
29. (b)	**30. (b)**	**31. (a)**	**32. (d)**
33. (c)			

CHAPTER 13

INDEX NUMBERS

1. CONCEPT AND DEFINITION OF INDEX NUMBERS

The concept of index number can be best understood through an illustration. Let us consider a situation of rising prices during the year 2018. In this context, we are faced with three basic questions. First, compared to which year have the prices risen during 2018? Second, how do we handle the situation when the prices of some goods rise more than the others? Thirdly, can prices of different goods be expressed in terms of any standard unit or different units are to be used to express prices of different goods and services, such that the price of milk is to be expressed in terms of rupees per liter, of cloth in terms of rupees per meter and of sweets in terms of rupees per kilogram. The study of Index Numbers answers all these questions: First, rise in prices during 2018 would be studied only with reference to some previous years like 2001 or 2004. Otherwise, the mere statement that prices during 2018 have tended to rise would make no sense. 2018 will be treated as the current year and 2001 or 2004 as the base year. Prices during the base year are taken as 100. Prices during the current year are related to the base year price. So that, we find out percentage change in price level over the years. Level of price is called the index of price. Since price of the base year is assumed to be 100, we can say that index of price of the base year is always 100. If base year is 2004 and the price index is 100, and if in the year 2018 prices are doubled, we shall conclude that the index of price in the year 2018 has shot up to 200 compared to 100 in the base year. As regards the second question (how do we handle the situation when prices of some goods change more than the others) the study of index numbers suggests to take average change. Thus, if the price of Potatoes has rises from 100 to 200 and the price of Onions has risen from 100 to 300, we shall consider average change only, viz. $\frac{200+300}{2} = 250$

Thus, it is the average index of prices for the various goods and services which is constructed for different years, and compared with the base year.

Third, as regards the problem of expressing the prices of various goods and services in some homogeneous units, the theory of Index Numbers suggests to consider only percentage change in prices of various goods and services. Once a change in price is expressed as a percentage change, the unit of the commodity (like litre of a milk, or meter of a cloth) loses its relevance.

Thus, what does the study of Index Numbers do? It helps us to find out percentage change in the values of different variables (may be prices of different goods or production of different commodities) over time with reference to some base year which happens to be the year of comparison. When various goods are studied simultaneously, the percentage change is taken as the average for all the goods.

Definition

In the words of Spiegel, "An Index Number is a statistical measure designed to show changes in a variable or group of related variables with respect to time, geographic location or other characteristics".

According to Croxton and Cowden, "Index Numbers are devices for measuring difference in the magnitude of a group of related variables".

2. FEATURES OR CHARACTERISTICS OF INDEX NUMBERS

Following are the three specific features or characteristics of index numbers:

(1) **Relative Changes:** Index numbers measure relative or percentage changes in the variable(s) over time. Index number of prices, for example, is not simply a statement of prices at different dates, it presents estimates of percentage changes in the prices over years with reference to some selected base year. If index of prices stands at 200 in 2018 compared to 100 in 2004-05 (the base year), it suggests that compared to the base year, prices have risen by 100 per cent.

(2) **Quantitative Expression:** Index numbers offer a precise measurement of the quantitative change in the concerned variable(s) over time. The index of prices, for example, will tell us that between the years 2017 and 2018, prices have risen by 7 percent, or that industrial production has declined by 3 per cent or that national income has risen just by 3.5 per cent during this period. The index numbers are not the qualitative statements like prices are rising or falling.

(3) **Averages:** Index numbers show changes in terms of averages. For example, when it is said that between the years 2017 and 2018, prices have risen by 7 per cent, it does not mean that prices of all goods and services have uniformly risen by 7 per cent; it only means that on an average there has been a 7 per cent rise in the prices of various goods and services. Even when prices of certain goods might have risen by more than 7 per cent and of certain others by less than 7 per cent.

3. DIFFICULTIES OR PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS

(1) **Purpose of Index Number:** There are various types of index numbers, constructed with different objectives. Before constructing an index number, one must define the objective. The construction of index number is significantly influenced by the objective or purpose of the study. Thus, for example, if the objective is to study the impact of change in the value of money on the consumers one should construct

consumers’ price index number. If we are to study the impact of change in the purchasing power of money on the producers, we shall construct index number on the basis of wholesale prices. Haberler has rightly pointed out that, ''Different index numbers are constructed to fulfil different objective and before setting to construct a particular index number, one must clearly define one’s object of study because, it is on the objective of the study, that the nature and format of the index number depends.”

(2) Selection of Base Year: Selection of Base Year is another problem in the construction of index number. Base year is the reference year. It is the year with which prices of the current year are compared. As far as possible, Base Year should be a normal year.

That is, it should be the one without much ups and downs. Otherwise, the index values would fail to capture the real change in the variable. The year 2004-05 is treated as base year in India, at present.

(3) Selection of Goods and Services: Having defined the objective, the problem is of the selection of goods or Services to be included in the index number. To construct the Consumers’ Price Index, for example, all commodities are not included. It is neither possible nor desirable to include all the goods and services produced in the country. We have to choose those goods and services which represent most of

Purpose of Constructing Index Number of

(i) Prices, and

(ii) Quantities

(i) Purpose of constructing index number of prices is to know the relative change or percentage change in the price level (made up of simple or weighted average of the prices of different goods and services) over time.

A rising general price level over time is a pointer towards inflation, while a falling general price level is a pointer towards deflation. Both inflation as well as deflation have notable consequences for an overall economic activity in the economy.

(ii) Purpose of constructing index number of quantity is to know relative change or percentage change in the quantum or volume of output of different goods and services over time. This reflects the level of economic activity in the economy and its different sectors. A rising index of quantity suggest a rising level of economic activity and vice versa.

others in the market. In other words, commodities selected should be such as are widely consumed, for example, rice, milk, ghee, cloth, etc. Larger the number of goods and services more representative is the index number.

(4) Selection of the Prices of the Goods and Services: Having selected the goods and services, the problem arises of prices to be selected. Broadly, in the construction of Price Index, the problem is whether to adopt retail prices or wholesale prices, controlled or open market prices. The choice would depend upon the objective or purpose of the study.

(5) Finding the Average Prices: In the construction of index number, base year value is assumed to be 100 and other values of different years are related to 100. Thus, if cloth price is Rs. 5 per meter in the base year and is found to be Rs. 10 per meter in the current year, the index of prices of cloth would be $\frac{10}{5} \times 100 = 200$ for the current year.

Likewise, price relatives for other commodities are worked out and average for these price relatives is determined and compared with the base year value of 100. It may be noted here that average of base year remains 100, but the average of the year under investigation may be more or less than 100. In case the average of the year under investigation is more than the average of the base year, it means that general price level has gone up. If it is less than the base year, it means that general price level has gone down. Generally, base year is indicated as 'O' and current year as T\ Price index is written as P_{01} and it is read as price index of year 1 in relation year 0.

(6) Selection of Weights: While constructing index number, weights are accorded to different commodities according to their relative significance. There are several methods of according weight, e.g., Fisher's method, Paasche's method, Laspeyre's method. While constructing weighted index number, one must justify his choice of weighting technique in accordance with the nature and objective of his study.

(7) Choice of Average: In finding out average values, different kinds of average may be used, geometric average, arithmetic average, etc. The choice of average significantly influences the results. Different kinds of averages may give different index number of a given change in price.

(8) Selection of Formula: Index numbers can be constructed with the help of many formulae, such as, Laspeyre's method, Paasche's method, Dorbish and Rowley's method, Fisher's method. One has to decide about the method to be used while constructing the index number.

4. ADVANTAGES OR USES OF INDEX NUMBERS

Some of the main advantages or uses of index numbers are as under.

(1) Measurement of Change in the Price Level or the Value of Money: Most important use of index numbers is that index numbers measure the value of money during different periods of time. We can use index numbers to know the impact of the change in the value of money on different sections of the society. Accordingly, devices or means can be worked out to correct inflationary or deflationary gaps in the system.

(2) Knowledge of the Change in Standard of Living: Index numbers help to ascertain the living standards of people. Money incomes may increase but if index numbers show a decrease in the value of money, living standards may even decline. Thus, index numbers indicate change in real income.

(3) Adjustments in Salaries and Allowances: Cost of living index is a useful guide to the Government and Private Enterprises to make necessary adjustments in salaries and allowances of the workers. Increase in the cost of living index suggests increase in salaries and allowances.

(4) Useful to Business Community: Price index numbers serve as a useful guide to the business community in their planning and decisions. Trend of the prices significantly influence their production decisions.

(5) Information Regarding Production: Index numbers of production shows whether the level of agricultural and industrial production in the economy is increasing or decreasing. Accordingly, agricultural and industrial development policies are formulated.

(6) Information Regarding Foreign Trade: Index of exports and imports provides useful information regarding foreign trade. Accordingly, export-import policies are for multi ted.

(7) Useful to Politicians: Politicians come to know of the real economic condition in the country on the basis of index numbers. They offer constructive criticism of government’s economic policies and give suggestions for economic reforms in the country.

Principal Limitations of Index Numbers

These are:

(i)there are no scientific techniques of according weightage to different items included in the index numbers.

(ii)weightage to different items is often influenced by personal bias.

(iii) owing to difference in the unit of currency as well as difference in the composition of production (and consumption) across different countries of the world, it is often very difficult to construct Index Numbers that facilitate international comparisons.

(8) Useful to Government: It is with the help of index numbers that the government determines its monetary and fiscal policies and takes concrete steps for the economic development of the country. In other words, with the help of index numbers government formulates appropriate policies to increase investment, output, income, employment, trade, price level, consumption, etc.

5. LIMITATIONS OF INDEX NUMBERS

In the construction of index numbers, there are some practical difficulties and theoretical limitations. The same are as under:

(1) Not Completely True: Index numbers are not fully true. For example, one can only make an estimate of change in the value of money with the help of index numbers. The index numbers simply indicate arithmetical tendency of the temporal changes in the variable.

(2) International Comparison not Possible: Different countries have different basis of index numbers. These do not help international comparisons.

(3) Difference of Time: With the passage of time, it is difficult to make comparisons of index numbers. With the changing times, man’s habits, tastes, etc., also undergo a change. Consequently, index numbers constructed on the basis of old consumption pattern cannot be compared with the index numbers constructed on the basis of new consumption pattern.

(4) **Limited Use:** Index numbers are prepared with certain specific objective. If they are used for another purpose they may lead to wrong conclusions. For example, index numbers prepared to know about the economic condition of the teachers cannot be used to know about the economic condition of the labourers.

(5) **Lack of Retail Price Index Numbers:** Most of the index numbers are prepared on the basis of wholesale prices. But in real life, retail prices are most relevant, but it is difficult to collect retail prices. Index numbers based on wholesale prices may be misleading. t

With regard to the limitations of index numbers, Coulbourn has rightly said, "In this changing world it is difficult to escape from the theoretical defects and in future, as far as we can see, it will not he possible, from theoretical point of view, to make use of the best method, of constructing the index number."

6. SIMPLE AND WEIGHTED INDEX NUMBERS

'Simple' and 'weighted' are the two broad categories of index numbers. Here is a brief description of these concepts.

Simple Index Numbers

These are the index numbers in which all items of the series are accorded equal weightage or importance. In case of a simple index of prices, for example, all goods and services are to be accorded equal weightage, no matter whether sale/purchase of certain goods is more than that of the others. It will be a simple average of the prices of different goods and services.

Weighted index Numbers

These are the index numbers in which different items of the series are accorded different weightage, depending upon their relative importance. It is not a simple average of prices of different goods and services, as in case of a simple price index. Instead, it is to be a weighted average of the prices of different goods. Thus, if the expenditure on rice is twice the expenditure on cloth, then in the construction of price index, price of rice may be accorded '2' as the weight compared to the weightage of T' to the price of cloth.

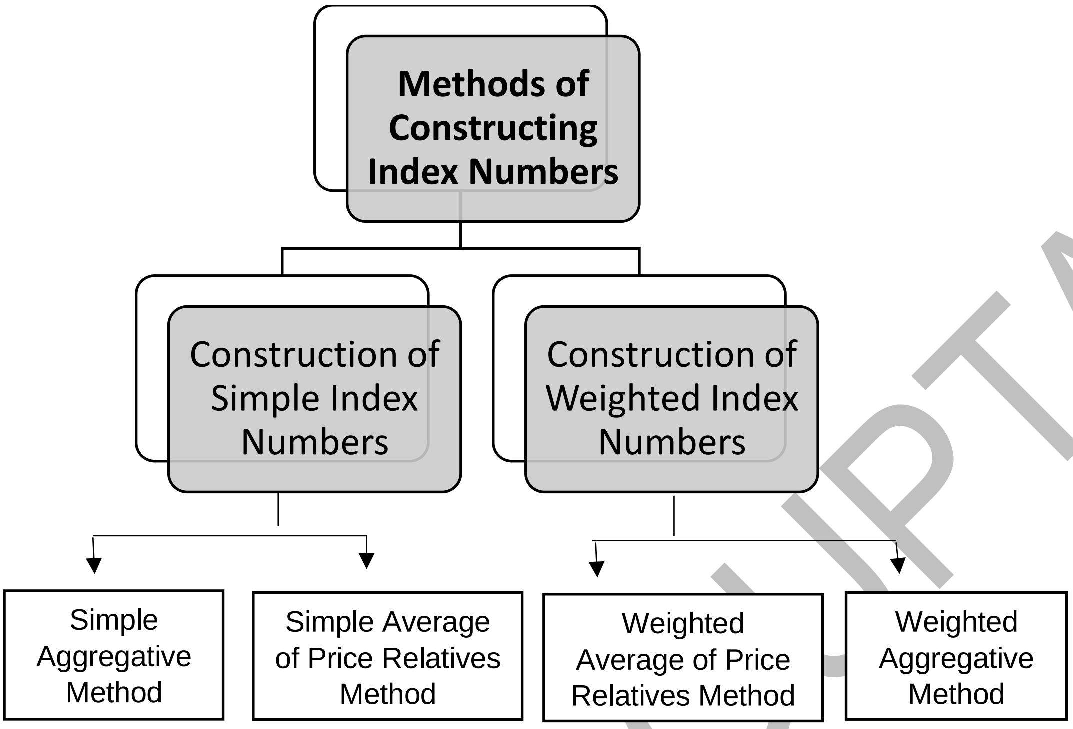
Though difficult to construct, weighted index numbers certainly offer a much more realistic view of the change over time compared to the simple index numbers.

The basic difference between Simple Index and Weighted Index

In the simple index, all items of the series are treated as of equal importance. In the weighted index, weights are accorded to different items depending on their relative importance.

7. METHODS OF CONSTRUCTING INDEX NUMBERS

The following chart shows the various methods of constructing index numbers (Simple as well as weighted):


Methods of Constructing Index Numbers
Construction of Simple Index Numbers
Construction of Weighted Index Numbers
Simple Aggregative Method
Simple Average of Price Relatives Method
Weighted Average of Price Relatives Method
Weighted Aggregative Method

Let us attempt a brief description of the various methods.

Base Year and its Characteristics

Base year is the year of comparison, also called reference year. It should bear the following characteristics:

(i) It should be a normal year, not showing wide fluctuations in the parameters related to the index number.

(ii)It should be a year for which reliable statistical data are available, so that comparison of the performance of the other years with the base year becomes meaningful-

(iii) It should not be a year too far from the period of study. Otherwise, relative change over time would not make much sense,

(iv) It should be neither very long nor very short period. Generally, it is not more than a year and not less than a month.

CONSTRUCTION OF SIMPLE INDEX NUMBERS

There are two methods of constructing simple index numbers:

(1) Simple Aggregative Method

In this method, aggregate of the prices of commodities in the current year are divided by the aggregate of their prices in the base year and multiplied by 100 to get index value for the current year. It is expressed by the following formula:

FORMULA

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

∑Po

Here, P_{01} = Price index of the current year.

ΣP_L = Sum of the prices of the commodities in the current year.

ΣP_0 = Sum of the prices of the commodities in the base year.

Current Year: Current year is the year for which average change is to be measured or index number is to be calculated.

Base Year: Base year is the year of reference from which we want to measure extent of change in the current year. The index number of base year is generally assumed to be 100.

Illustration.

Given the following data and assuming 2004 as the base year, find out index value of the prices of different commodities for the year 2018.

Commodity	A	B	C	D	E
Price in 2004 (Rs.)	50	40	10	5	2
Price in 2018 (Rs.)	80	60	20	10	6

Solution:

Construction of a Simple Index Number— Simple Aggregative Method

Commodity	2004 Price (Rs.) (P_0)	2018 Price (Rs.) (P_1)
A	50	80
B	40	60
C	10	20
D	5	10
E	2	6

Total	$\Sigma P_0 = 107$	$\Sigma P_1 = 176$

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

$$= \frac{176}{107} \times 100$$

= 164.49

Price Index = 164.49.

(2) Simple Average of Price Relatives Method

According to this method, we first find out price relatives for each commodity and then take simple average of all the price relatives.

What is Price Relative?

A price relative is the percentage ratio of the value of a variable in the current year to its value in the base year. In other words, a price relatives is a percentage ratio between price of a commodity in the current year and that in the base year.

$$\text{Price Relatives, } P_{01} = \frac{\text{Current Year Price } (P_1)}{\text{Base Year Price } (P_0)} \times 100$$

We can find out price index number of the current year by using the following formula.

FORMULA

$$P_{01} = \frac{\Sigma(\frac{P_1}{P_0} \times 100)}{N}$$

An Important Caution

Prices for different commodities are expressed with reference to different units of measurement. Price of steel, for example, is expressed as rupee per kg, while the price of milk is expressed as rupee per liter, and the price of doth is expressed as rupee per meter. Simple aggregate method of index numbers cannot be used for commodities with different units of measurement. It can be used only for those commodities which have a common unit of measurement.

(Here, $\frac{P_1}{P_0} \times 100\times 100$ = Price relatives; N = Number of goods; P, = Current year’s value; P_0 = Base year’s value.)

Illustration.

Given the following data and using the Price Relatives Method, construct an index number for the year 2018 in relation to 2004 prices.

Commodity	**Wheat**	**Ghee**	**Milk**	**Rice**	**Sugar**

2004 Price (Rs.)	100 (per qt.)	8 (per kg)	2 (per l)	200 (per qt.)	1 (per kg)
2018 Price (Rs.)	200	40	16	800	6

Notation of 0 and 1

Often the notation of 0 stands for the base year, and 1 stands for the current year.

Solution:

Construction of a Simple Index Number— Simple Average of Price Relative Method

Commodity	**Base Year 2004 Price (P_0)**	**2018 Price (P_1)**	**Price Relatives of 2018 in relation to 2004** $\left(\frac{P_1}{P_0} \times 100\right)$
Wheat	100 (per qt)	200 (per qt)	$\frac{200}{100} \times 100 = 200$
Ghee	8 (per kg)	40 (per kg)	$\frac{40}{8} \times 100 = 500$
Milk	2 (per l)	16 (per l)	$\frac{16}{2} \times 100 = 800$
Rice	200 (per qt)	800 (per qt)	$\frac{800}{200} \times 100 = 400$
Sugar	1 (per kg)	6 (per kg)	$\frac{6}{1} \times 100 = 600$
N = 5			$\Sigma\left(\frac{P_1}{P_0} \times 100\right) = 2{,}500$

$$P_{01} = \frac{\Sigma(\frac{P_1}{P_0} \times 100)}{N}$$

$$= \frac{2{,}500}{5} = 500$$

Price Index = 500.

CONSTRUCTION OF WEIGHTED INDEX NUMBERS

There are two methods of constructing weighted index numbers, as discussed below:

(1) Weighted Average of Price Relatives Method

Not all Commodities are Ever Included

Not all commodities are ever included in the construction of an Index Number. Only a sample of commodities is taken which represents characteristics of the entire group of commodities under study.

According to this method, weighted sum of the price relatives is divided by the sum total of the weights. In this method, goods are given weight according to their quantity. Thus,

FORMULA

$$P_{01} = \frac{\sum RW}{\sum W}$$

(Here, P_{01} = Index number for the current year in relation to the base year; W = Weight; R = Price relative.)

Illustration.

Given the following data and using Weighted Average Price Relative Method, construct index number for 2018 based on 2004 prices.

Goods	Weight	2004 Price (Rs.)	2018 Price (Rs.)
Wheat	40	100 (per qt)	200 (per qt)
Rice	30	200 (per qt)	800 (per qt)
Miik	15	2 (per l)	16 (per l)
Ghee	10	8 (per kg)	40 (per kg)
Sugar	5	1 (per kg)	6 (per kg)

Solution:

Goods	Weight (W)	2004 Price (P_0)	2018 Price (P_1)	$R = \frac{P_1}{P_0} \times 100$	RW
Wheat	40	Rs. 100 per qt	Rs. 200 per qt	$\frac{200}{100} \times 100 = 200$	200 × 40 = 8,000

Rice	30	Rs. 200 per qt	Rs. 800 per qt	$\frac{800}{200} \times 100 = 400$	400 × 30 = 12,000
Milk	15	Rs. 2 per l	Rs. 16 per l	$\frac{16}{2} \times 100 = 800$	800 × 15 = 12,000
Ghee	10	Rs. 8 per kg	Rs. 40 per kg	$\frac{40}{8} \times 100 = 500$	500 × 10 = 5,000
Sugar	5	Rs. 1 per kg	Rs. 6 per kg	$\frac{6}{1} \times 100 = 600$	600 × 5 = 3,000
Total	∑W = 100				∑RW = 40,000

$$P_{01} = \frac{\Sigma RW}{\Sigma W} = \frac{40{,}000}{100} = 400$$

Price Index = 400.

(2) Weighted Aggregative Method

Under this method different goods are accorded weight according to the quantity bought. Economists have different views in this respect. Should the weight be given (i) on the basis of the quantity bought in the current year or (ii) on the basis of the quantity bought in the base year or (iii) on the basis of the quantities bought in both the years? Different economists have, therefore, suggested different techniques of weighting. Some of the well-known methods are as under:

What is the basic difference between Laspeyre's and Paasche's methods of construction of Weighted index Numbers?

Laspeyre’s uses base year quantities as the weights of different items. Paasche’s on the other hand, uses current year quantities as weights.

(i) Laspeyre’s Method: Laspeyre’s uses base year quantities (q_0) as weights of different items. His formula for estimating Index values is:

FORMULA

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

(ii) Paasche’s Method: Paasche’s on the other hand uses current year's quantities (q_t) as weight. His formula to construct the Index value is:

FORMULA

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

(iii) Fisher’s Method: Fisher has combined the techniques of Laspeyre’s and Paasche’s method. He used both base year as well as current year quantities (q_0, q,) as weight. His formula to construct Index Number is:

FORMULA

$$P_{01}=\sqrt{\frac{\sum \mathrm{P}_1\mathrm{Q}_0}{\sum \mathrm{P}_0\mathrm{Q}_0}\times\frac{\sum \mathrm{P}_1\mathrm{Q}_1}{\sum \mathrm{P}_0\mathrm{Q}_1}}\times 100$$

Fisher’s method is treated as Ideal Formula.

FISHER'S INDEX NUMBER AS AN IDEAL METHOD

The choice of method for the construction of an index number will depend upon the object with which a particular index number is constructed. Many formulae may be used for the construction of index numbers but all may not be suitable for the specific purpose in hand. Some of the important formulae do not conform to certain appropriate test of consistent behaviour; it implies that these formulae give us biased results.

However, Fisher’s Method is considered as an ideal method for constructing index numbers:

$$P_{01}=\sqrt{\frac{\sum P_1q_0}{\sum P_0q_0}\times\frac{\sum p_1q_1}{\sum p_0q_1}}\times 100$$

Fisher’s method is considered as idea because

(i) It is based on variable weights.

(ii) It takes into consideration the price and quantities of both the base year and current year.

(iii) It is based on geometric mean (GM) which is regarded as the best mean for calculating index number.

(iv) Fisher’s index number satisfies both the **Time Reversal Test and Factor Reversal Test.** The time reversal test implies that the formula for calculating an Index Number should be such that will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base. Time Reversal means that if we change base year to current year and vice versa then the product of two indexes should be equal to unity. Thus, an index number should work both ways, i.e., forward as well as backward. On the other hand, Factor Reversal Test implies, just as our formula should permit the interchange of two items without giving inconsistent results, so it ought to permit interchange of prices and quantities without giving inconsistent results, i.e., the two results multiplied together should be equal to value ratio.

Illustration.

Construct index numbers of prices of the items in the year 2018 from the following data by:

(i) Laspeyre’s Method,

(ii) Paasche’s Method, and

(iii) Fisher’s Method.

Items	2004 (Base Year)		2018 (Current Year)	
	Price	Quantity	Price	Quantity
A	10	10	20	25
B	35	3	40	10
C	30	5	20	15
D	10	20	8	20
E	40	2	40	5

Solution:

Construction of Price Index Numbers

Items	Base Year (2004)		Current Year (2018)		p_0q_0	p_0q_1	P_1q_0	P_1q_1
	Price (p_0)	Quantity (q_0)	Price (P_1)	Quantity (q_1)				
A	10	10	20	25	100	250	200	500
B	35	3	40	10	105	350	120	400
C	30	5	20	15	150	450	100	300
D	10	20	8	20	200	200	160	160
E	40	9	40	5	80	200	80	200

					$\sum p_0q_0 = 635$	$\sum p_0\, q_1 =$ 1,450	$\sum p_1q_0=$ 660	$\sum p_1q_1 =$ 1,560

(i) Laspeyre’s Method: $P_{01} = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$

$$= \frac{660}{635} \times 100 = 103.94$$

(ii) Paasche’s Method: $P_{01} = \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100$

$$= \frac{1{,}560}{1{,}450} \times 100 = 107.59$$

(iii, Fisher’s Method: $P_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$

$$= \sqrt{\frac{660}{635} \times \frac{1{,}560}{1{,}450}} \times 100$$

$$= \sqrt{1.03 \times 1.07} \times 100$$

$$= \sqrt{1.1021} \times 100$$

$$= 1.05 \times 100 = 105$$

Illustration.

Find out index value from the following data using (i) Laspeyre’s Method, (ii) Paasche’s Method, and (iii) Fisher’s Method.

Items	Base Year Quantity	Base Year Price (Rs.)	Current Year Quantity	Current Year Price (Rs.)
A	6	10	8	12
B	4	15	5	20
C	5	8	3	16
D	3	9	6	1

Solution:

Items	q_0	p_0	q_1	p_1	p_1q_0	P_0q_0	p_1q_1	p_0q_1

A	6	10	8	12	72	60	96	80
B	4	15	5	20	80	60	100	75
C	5	8	3	16	80	40	48	24
D	3	9	6	1	3	27 4	6	54
					$\sum p_1q_0 = {}^{235}$	$\sum P_0q_0 = {}^{187}$	$\sum p_1q_1 = 250$	$\sum p_0q_1 = {}^{233}$

(i) Laspeyre’s Method: $P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

$$= \frac{235}{187} \times 100 = 125.67$$

(ii) Paasche’s Method: $P_{01} = \frac{\Sigma \mathrm{p}_1 \mathrm{q}_1}{\Sigma \mathrm{P}_0 \mathrm{q}_1} \times 100$

$$= \frac{250}{233} \times 100 = 107.29$$

(iii) Fisher’s Method, $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$

$$= \sqrt{\frac{235}{187} \times \frac{250}{233}} \times 100$$

$$= \sqrt{1.2567 \times 1.0729} \times 100$$

$$= \sqrt{1.3483} \times 100$$

= 1.1612 x 100 = 116.12

Illustration.

From the following data construct Fisher’s Ideal Index:

Commodity	Base Year Price	Base Year Quantity	Current Year Price	Current Year Quantity
A	5	50	7	60
B	6	15	8	10

C	8	8	11	12
D	7	20	10	15

Solution:

Commodity	q_0	p_0	q_1	p_1	p_1q_0	P_0q_0	p_1q_1	p_0q_1
A	5	50	7	60	350	250	420	300
B	6	15	8	10	120	90	80	60
C	8	8	11	12	88	64	132	96
D	7	20	10	15	200	140	150	105
					$\sum p_1q_0 = 758$	$\sum P_0q_0 = {}^{544}$	$\sum p_1q_1 = 782$	$\sum p_0q_1 = 561$

$$P_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$$

$$= \sqrt{\frac{758}{544} \times \frac{782}{561}} \times 100$$

$$= \sqrt{1.9423} \times 100$$

= 139.37

8. CONSUMER PRICE INDEX OR COST OF LIVING INDEX NUMBER

So far we have been focusing on general price indices. But these indices do not precisely explain how the change in general price level affects the cost of living of rife various classes of society. This is because different classes of people consume different goods and services. Accordingly, the change in prices affect them differently. To know the effects of changing prices on the living of different classes of society, we need a special type of price index, called **Consumer Price Index or Cost of Living Index Number.**

The consumer price index is the index number which measures the average change in prices paid by the specific class of consumers for goods and services consumed by them in the current year in comparison with base year. Change in the price level affects the cost of living of the concerned class of consumers. Accordingly, consumer price indices are also called **cost of living indices.**

In India, the consumer price indices are mainly constructed for the following consumer groups:

(i) Industrial Workers (IW)

(ii) Urban-Non-Manual Employees (UNME)

(iii) Agricultural Labourers (AL).

Construction of Consumer Price index

Construction of the consumer price index number involves the following steps:

(1) Selection of the Consumer Class: Consumers are classified into various classes like industrial labour, government employees, agricultural labour, teachers, etc. We should select the class of consumer according to the requirement of our study.

(2) Information about the Family Budget: After the decision about the group, a sample of persons should be selected from the group and following information about their family budgets should be obtained:

(i) Commodities which they consume.

(ii) Quantity of consumption.

(iii) Prices of the concerned goods and services.

(iv) Money spent on these goods and services.

The items which are generally consumed may be classified in groups like (i) Food, (ii) Clothing, (iii) Fuel and light,

(iv) House rent, (v) Education, health and sanitation, (vi) Miscellaneous.

(3) Choice of Base Year: The base year should be the ^ear of economic stability. It should not be too distant from the current year. Presently, 2011-12 is being used as base year.

(4) Information about Prices: The retail prices of selected items/ commodities should be collected from the region where the group of selected person lives and makes the purchases.

(5) Weightage: The importance of various items for different classes of people is different. Therefore, the selected items should be given weights according to their relative importance. As discussed earlier, there are two ways of according weights:

(i) Quantity Weights: These weights are given in proportion to the quantities consumed in the base period.

(ii) Expenditure Weights: These weights are given in proportion to the total expenditure on the items consumed in the base period.

Methods of Constructing Consumer Price index (CPI)

Corresponding to the two methods of assigning weights to different commodities, there are two methods of the construction of Consumer’s Price Index.

(1) Aggregative Expenditure Method

(2) Family Budget Method.

(1) Aggregative Expenditure Method: This method is similar to

the Laspeyre’s method (already discussed).

The following formula is used in this method.

FORMULA

Consumer Price Index (CPI) = $\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

where, p_1 = Price of the commodities in the current year.

p_0 = Price of the commodities in the base year.

Thus,

q_0 = Quantity consumed in base year.

$\sum p_1 q_0$ shows aggregate expenditure in the current year.

$\sum p_0 q_0$ shows aggregate expenditure in the base year.

$$P_{01} = \frac{\sum p_1 q_0}{\Sigma P_0 q_0} \times 100$$

(2) Family Budget Method: The following formula is used in this method to find Consumer’s Price Index.

FORMULA

Consumer’s Price Index = $\frac{\Sigma RW}{\Sigma W}$.

where, R = Current year’s price relative of various items. W = Weights of various items.

Calculation of CPI by this method involves the following steps:

(i) The current year’s price (pj) of each commodity is divided by the base year’s price (p_0) of the respective commodities. Then the resultant is multiplied by 100. These are called price relatives of the current year.

Therefore,

Price Relatives of the Current year $\frac{\text{Price of the Current Year}}{\text{Price of the Base Year}} \times 100$

$$R = \frac{P_1}{P_0} \times 100$$

(ii) Aggregate expenditure on each item is considered as the weight of the item. Hence, the weight (W) of a commodity is calculated by multiplying the price (p) of a commodity in the base year with the quantity (q_0) of the commodity consumed in the base year.

(iii) The price relative (R) of each item is multiplied with their respective weights (W = p_0q_0)- These products are added to find ∑RW.

(iv) The sum of weights ∑W or $\sum p_0q_0$ is calculated.

(v) These values are substituted in the following formula to find the consumer’s price index.

Consumer Price Index

$$\frac{\text{Sum of Products of the Price relative with weights}}{\text{Sum of Weights}}$$

Thus:

$$P_{01} = \frac{\sum RW}{\sum W}$$

Illustration.

Find the Consumer Price Index or the Cost of Living Index Number for the current year from the following data by (i) Aggregate Expenditure Method, and (ii) Family Budget Method.

Articles	Quantity Consumed in Base Year	Price in Base Year	Price in Current Year
Rice	5 qt	24 per qt	30 per qt
Wheat	1 qt	16 per qt	20 per qt
Pulses	2 qt	12 per qt	18 per qt
Ghee	4 kg	5 per kg	6.25 per kg
Oil	5 l	4 per l	5 per l
Clothing	40 meters	1 per meter	1.50 per meter
Firewood	10 qt	2 per qt	2.50 per qt
House Rent	1 house	20 per house	25 per house

Solution:

(i) Aggregative Expenditure Method

From the given data we derive the following table:

Articles	**Quantity Consumed in Base Year** **(q_0)**	**Price in Base Year** **(p_0)**	**Price in Current Year** **(p_0)**	**Aggregate Expenditure in Base Year** **($\rho_0 q_0$)**	**Aggregate Expenditure in Current Year ($p_1 q_0$)**
Rice	5 qt	24	30	120	150
Wheat	qt	16	20	16	20
Pulses	2 qt	12	18	24	36
Ghee	4 kg	5	6.25	20	25
Oil	5 l	4	5	20	25
Clothing	40 m	1	1.50	40	60
Firewood	10 qt	2	2.50	20	25
House Rent	1 house	20	25	20	25
				$\sum \rho_0 q_0 = {}^{280}$	$\sum p_1 q_0 = 366$

Index Number for Current Year = $\frac{\Sigma p_1 q_0}{\Sigma P_0 q_0} \times 100$

$$= \frac{366}{280} \times 100 = 130.71$$

(ii) Family Budget Method

Articles	**Quantity Consumed in Base Year** **(q_0)**	**Price in Base Year** **(p_0)**	**Price in Current Year** **(p_1)**	**Price Relative for Current Year** $\left(\boldsymbol{R} = \frac{\boldsymbol{P_1}}{\boldsymbol{p_0}} \times \mathbf{100}\right)$	**Weights** **(Value consumed in Base Year)** **($W = p_0 q_0$)**	**Product of Price Relative and Weight (RW)**
Rice	5 qt	24	30	125	120	15,000

Wheat	1 qt	16	20	125	16	2,000
Pulses	2 qt	12	18	150	24	3,600
Ghee	4 kg	5	6.25	125	20	2,500
Oil	5 l	4	5	125	20	2,500
Clothing	40 meters	1	1.50	150	40	6,000
Firewood	10 qt	9	2.50	125*	20	2,500
House Rent	1 house	20	25	125	20	2,500
					∑W = 280	∑RW=36,600

Now CPI for Current Year = $\frac{\Sigma^{RW}}{\Sigma^{W}} = \frac{36,600}{280} = 130.71$

Illustration.

Calculate the Cost of Living Index from the following data:

Items	Quantity Consumed in the given Year	Price per unit in Rs.	
		Base Year	Given Year
Rice	$2\frac{1}{2}$qt × 12	12	25
Pulses	3 kg x 12	0.4	0.6
Oil	2 l X12	1.5	2.2
Clothing	6 meters X 12	0.75	10
Housing	12 months	20 per month	30 per month
Miscellaneous	Expenditure of 12 months	10 per month	15 per month

Solution:

Items	Quantity Consumed in the given Year (q_1)	Price		Aggregate Expenditure in the Base Year (p_0p_1)	Aggregate Expenditure in the Current Year (p_1q_1)
		Base Year (p_0)	Current Year (p_1)		
Rice	30 qt	12	25	360	750
Pulses	36 kg	0.4	0.6	14.4	21.6
Oil	24/	1.5	2.2	36	52.8
Clothing	72 meters	0.75	10	54	720
Housing	12 months	20 per month	30 per month	240	360
Miscellaneous	12 months expenditure	10 per month	15 per month	120	180
				$\sum p_0p_1 =$ 824.4	$\Sigma\ p_1q_1 =$ 2,084.4

Cost of Living Index = $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$

$$= \frac{2{,}084.4}{824.4} \times 100 = 252.84$$

Illustration.

Construct Cost of Living Index for 2018 based on 2004 from the following data:

Group	Food	Housing	Clothing	Fuel and Light	Miscellaneous
Group Index Number for 2018 (Based on 2004)	122	140	112	116	106
Weights	32	10	10	6	42

Solution:

Group	Group Index Number (R)	Weights (W)	Weighted Relative (RW)
Food	122	32	3,904
Housing	140	10	1,400
Clothing	112	10	1,120
Fuel and light	116	6	696
Miscellaneous	106	42	4,452
		∑W=100	∑RW = 11,572

Thus, Cost of Living Index = $\frac{\sum RW}{\Sigma W} = \frac{11,572}{100} = 115.72$

Uses

(i) Formulation of Price Policy.

(ii) Wage Adjustment.

(iii) Measurement of Real Value.

(iv) Analysis of Markets.

(v) National Income Deflator.

Importance of the Consumer Price Index or Cost of Living Index

(1) Formulation of Price Policy: The consumer price indices are used by government to frame policies on prices. On the basis of these indices government decides whether the prices are to be controlled, dual price policy should be adopted or public distribution system is to be introduced, etc. Also, government policies like rent control and taxation, general economic and fiscal policies etc. are framed on the basis of the consumer price index numbers to a large extent.

(2) Wage Adjustment: Cost of living index numbers are used as basis for the wage adjustments. The rates of dearness allowances are decided by the government on the basis of these indices. These indices are also used for wage contracts and wage agreements of the workers.

(3) Measurement of Real Value: These index numbers are used to measure the real value of the rupee or its purchasing power and real income (or revenue), etc.

(4) Analysis of Markets: The consumer price indices are also used for the analysis of the market of specific commodities for their demand and supply.

(5) National Income Deflator: These indices are also used as deflators of national income. Accordingly, real change in national income is estimated.

Difficulties in the Construction of Consumer Price Index

There are many difficulties in the construction of consumer price index. These are as follows:

(1) Difference in the Standards of Living: Consumption pattern of different classes of consumers are different and therefore, have different living standards. Thus, there cannot be one consumer price index number for different classes of society.

(2) **Difference in Prices:** The price indices are constructed on the basis of retail prices. But the retail prices vary from place to place and even at the same place from shop to shop. As such it is very difficult to find a representative price for the calculation of consumer price index.

(3) Difference in the Proportion of Expenditure: All the members of any particular group do not spend on various items of consumption in same ratio and even one person does not spend on various commodities in the same ratio at two different periods of time. A consumer's purchase ratio depends upon his/her taste, habits, etc. Accordingly, it is difficult to construct a cost of living index that truly reflects a change in the cost of living over time.

9. WHOLESALE PRICE INDEX (WPI)

The Wholesale Price Index (WPI) measures the relative changes in the prices of commodities traded in the wholesale markets. In India, the wholesale price index numbers are constructed on weekly basis. The year 2011-12 is being used as the base year.

Commodity Group and Weightage of Wholesale Price Index

In India, all the commodities have been classified in the following three groups:

Commodity Group	Name of Commodities	Weightage
(i) Primary Articles	These include 98 commodities like Rice, Fruits, Pulses, Vegetables and Non-food articles like Cotton, Jute, Metals.	22.02
(ii) Fuel, Power, Light and Lubricants	These include 19 items like Coal, Petroleum Products, Electricity, LPG.	14.23
(iii) Manufacturing	It includes 318 items like Textiles, Sugar, Paper, Machinery, Chemicals, Fertilizers, Leather, etc.	63.75

The basic difference of purpose behind the Consumer Price Index Number and the Wholesale Price Index Number

In case of consumer price index number, the basic purpose is to know cost of living of a specified group of consumers in the society. In case of wholesale index number, the basic purpose is to assess the situations of overall demand and supply in the economy. Rising prices indicate a situation of excess demand, while falling prices suggest a situation of deficient demand. Wholesale price index focuses on the rate of inflation in the economy.

Producer Price Index

As in many countries, in India also, efforts are afoot to shift from WPI (Wholesale Price Index) to Producer Price Index. Producer-Price refers to the basic price including taxes, trade margins and transport cost. Producer Price Index is expected to offer better insights into the analysis of price trends in the country.

Uses of Wholesale Price Index

(1) **Forecasting Demand and Supply:** The wholesale price indices are often used to forecast demand and supply situation in the economy. An increase in wholesale price index is an indication of excess demand. It is a situation in which demand is greater than supply. On the other hand, a decrease in wholesale price index implies deficient demand. It is a situation in which demand is less than supply.

(2) **Estimation of Monetary Value and Real Value:** The wholesale price index can be used to estimate the monetary value and real value of aggregates like national income and expenditure. Monetary value is the value estimated at current year prices. Real value is the value estimated at base year prices or at constant prices. The monetary aggregate can be converted into real aggregate by applying the following formula:

Real Aggregate of the Current Year = Monetary Aggregate of the Current Year

$$\frac{\text{Price Index of Base Year}}{\text{Price Index of Current Year}}$$

(3) **Indicator of Rate of Inflation**: The wholesale price index is also applied to calculate the rate of inflation in a country. It refers to the rate at which prices tend to increase over time.

What is Rate of Inflation?

Wholesale price index is prepared for every week. If for week 1, wholesale price index is A, and for week 2, the wholesale price index is A_2, then the rate of inflation between week 1 and week 2 would be estimated as under;

$$\frac{A_2 - A_1}{A_1} \times 100$$

Illustration.

If wholesale price index for week 1 = 200 and for week 2 = 250, then

Rate of inflation = $\frac{250-200}{200} \times 100$

$$= \frac{50}{1200} \times 100 = 25\%$$

Annual rate of inflation is estimated by considering average of the wholesale price index for all weeks of the year.

10. INDEX NUMBER OF INDUSTRIAL PRODUCTION

Index number of industrial production is that index which measures the relative increase or decrease in the level of industrial output in a country in comparison to the level of production in the base year. In India, the base year for the current series is 2011-12. These index numbers tell us about the changes in the quantum of production. These index numbers are useful in estimating the growth of industrial production in the economy.

Construction of Index Number of Industrial Production

Construction of the index number of industrial production involves the following steps:

(1) Classification of Industries: To construct index number of industrial production the industries are classified into following groups:

(i) Mining, (ii) Manufacturing, and (iii) Electricity.

(2) **Statistics or Data Related to Industrial Production:** The data relating to the production of the above-mentioned industries are collected either monthly, quarterly or yearly.

(3) **Weightage:** Weights are given on the basis of the relative importance of different industries. The weights are based on the values of net output of different industries, and their contribution to national income.

In India, the following weightage is given to different groups at present:

Group	Weightage
(1) Mining	10.47
(2) Manufacturing	79.36
(3) Electricity	10.17
Total	100.00

Index Number of Industrial Production is calculated by using the following formula:

FORMULA

Index Number of Industrial Production = $\frac{\Sigma(\frac{q_1}{q_0})W}{\Sigma W} \times 100$

where, q_1 = Level of production in the current year.

Q_2 = Level of production in the base year.

W = Weight or relative importance of industrial output.

Illustration.

Construct index number of industrial production from the following data:

Industry	**Output**			**Units**
	Base Year	**Current Year**	**Weights**	
Manufacturing Production	122	300	85	Mill. Tonnes
Electrical Products	203	400	5	Th. Nos.
Mining	65	87	10	Mill. Tonnes

Solution:

From the given data we have

	Output		**Relative Value**	**Weights**	**Weighted**
Industry	**Base Year** (q_0)	**Current Year** (q_1)	$\left(R = \frac{q_1}{q_0} \times 100\right)$	**(W)**	**Relative (RW)**
Manufacturing Production	122	300	$\frac{300}{122} \times 100 = 245.90$	85	20,901.50
Electrical Products	203	400	$\frac{400}{203} \times 100 = 197$	5	985.00
Mining	65	87	$\frac{87}{65} \times 100 = 134$	10	1,340.00
				∑W = 100	∑RW=23,226.50

Index Number of Industrial Production

$$= \frac{\sum RW}{\sum W} = \frac{23{,}226.50}{100} = 232.27(\text{approx.})$$

11. INFLATION AND INDEX NUMBERS

Inflation refers to a situation of rise in the general price level in a country over a long period of time. Often, inflation is measured in terms of wholesale price index. A consistent rise in the wholesale price index over time implies a situation of inflation. Example: If wholesale price index rises from 100 in 2011-12 (base year) to 150 in 2018, and if the increase in price has almost been consistent over time (like, every year the general price level has been rising by 5-10%), it would be deemed as a situation of inflation. Implying a continuous erosion in the value of money or purchasing power of money. Value of money (or purchasing power of money) reduces to half if wholesale price index rises by 100%. Obviously, double the price level, half the purchasing power of a rupee. If money income of the people remains constant, 10% increase in the price level implies a 10% decrease in purchasing power of the people.

Inflation causes erosion of purchasing power of the people, if their money income remains constant. Accordingly, often we find workers pestering the government for dearness allowance (DA) to compensate for the loss of purchasing power during periods of inflation.

Fall in the Rate of Inflation does not Imply Fall in the Price Level

Students here must appreciate the distinction between inflation and the rate of inflation. Inflation is measured as a percentage increase in the general price level, like from 100 in 2017 to 115 in 2018, implying a 15% increase in the price level during the period of one year. While prices have tended to increase during the year, these may not be rising at the same pace (or the same rate) every week within the year. Relative change in the price index from week to week measures the rate of inflation.

Rate of Inflation = $\frac{A_2 - A_1}{A_1} \times 100$

where, Aj = Wholesale price index for week 1, and Ag = Wholesale price index for week 2. Thus, within the year, rate of inflation may increase or decrease. It only implies the increase or decrease in the pace of inflation or increase/decrease in the speed of inflation not a fall in the price level. To illustrate, in week 1 the rate of inflation may be 5% while in week 2, the rate of inflation may be 4%. Fall from 5% to 4% should not be interpreted as a fall in the price level. It only implies a fall in the speed at which prices tend to rise.

Inflation is measured in terms of changes in the wholesale price index, based on weekly quotations of wholesale prices.

increase in the wholesale price index over a long period of time implies a situation of inflation. It causes erosion of purchasing power of money.

In a situation of inflation, trade unions often pester for higher money wages to compensate for the fall in the purchasing power of a rupee.

Sensex

Sense* is the index showing changes in the Indian stock market It is a shirt form of Bombay Stock Exchange Sensitive Index. It is constructed with 1978-79 σs the reference

year or the base year. It consists of 30 stocks of the leading companies in the country. Changes in these stocks are expected to represent changes in the entire stock market. Rise in sensex implies an overall increase in share prices. This shows increase in expected earnings from investments in the stock market. This also implies robust performance of the principal industries in the economy.

Multiple Choice Questions

Select the correct alternative:

1. Whose formula is considered ideal for the construction of index number?

(a) Paasche's formula

(b) Laspeyer's formula

(c) Fisher's formula

(d) None of these

2. If the index of prices is estimated to be 112 in 2018, it means that in comparison to the base year, prices in 2018 are higher by:

(a) 12%

(b) 12 × 2 = 24%

(c) 112%

(d) none of these

3. Which of the following formulae is propounded by Fisher?

(a) $P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$

(b) $P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

(c) $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$

(d) None of these

4. In notation P_{01}, 1 stand for:

(a) current year

(b) reference year

(c) both (a) and (b)

(d) none of these

5. Which of the following equations is correct?

(a) $P_{01} = \frac{\Sigma RW}{\Sigma W} \times 100$

(b) $P_{01} = \frac{\Sigma RW}{\Sigma W}$

$(c) P_{01} = \frac{\Sigma W}{\sum RW} \times 100$

$(d) P_{01} = \frac{\Sigma W}{\Sigma RW}$

6. Base year is also known as:

(a) current year

(b) reference year

(c) periodic year

(d) both (a) and (c)

7. Price Relatives = $\frac{\text{Current Year Price}}{?} \times 100$

(a) Reference year price

(b)Periodic year price

(c) Base year price

(d)Both (a) and (c)

8. $P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$ is the formula of:

(a) Laspeyre's Method

(b) Paasche's Method

(c) Fisher's Method

(d) none of these

9. Fisher's method of calculating index numbers is based on:

(a) arithmetic mean

(b) harmonic mean

(c) geometric mean

(d) none of these

10. Fisher's index number is considered ideal because:

(a) it is based on variable weights

(b) it satisfies Time Reversal Test

(c) it satisfies Factor Reversal Test

(d) all of these

11. Consumer Price Index is also known as:

(a) Industrial Production Index

(b) Cost of Living Index

(c) Wholesale Price Index

(d) none of these

12. Rate of inflation is equal to:

(a) $\frac{A_1}{A_2 + A_1} \times 100$

(b) $\frac{A_2+A_1}{A_1} \times 100$ [A = Price index]

(c) $\frac{A_1}{A_2 - A_1} \times 100$

(d) $\frac{A_2 - A_1}{A_1} \times 100$

13. The Paasche's index number is based on:

(a) Base year quantities

(b) Current year quantities

(c) Average of current and base years

(d) None of the above

14. Index number for the base period is always taken as:

(a) 100

(b) 50

(c) 1

(d) 200

15. Fisher's Ideal index is the:

(a) Mean of Laspeyre's and Paasche's indexes

(b) Median of Laspeyre's and Paasche's indexes

(c) Geometric mean of Laspeyre's and Paasche's indexes

(d) None of the above

16. The aggregate index formula using base period quantities is known as:

(a) Laspeyre's index

(b) Fisher's Ideal index

(c) Bowley's index

(d) Paasche's index

17. The index used to measure changes in total money value is called:

(a) Price index

(b) Quantity index

(c) Value index

(d) None of the above

18. We use price index numbers:

(a) To measure and compare prices

(b) To compare prices

(c) To measure prices

(d) None of these

19. The best average for constructing an index number is:

(a) Harmonic Mean

(b) Arithmetic Mean

(c) Geometric Mean

(d) None of these

Answers

Multiple Choice Questions

1.(c)	**2. (a)**	**3. (c)**	**4. (a)**	**5. (b)**	**6. (b)**
7.(d)	**8. (a)**	**9. (c)**	**10. (d)**	**11. (b)**	**12. (d)**
13.(b)	**14. (a)**	**15. (c)**	**16. (a)**	**17. (c)**	**18. (a)**
19. (c)					

LIST OF FORMULAS

MEANING OF CENTRAL TENDENCY — MEAN

1. SIMPLE MEAN **Individual Series** **Direct Method**	$(\bar{X}) = \frac{\sum X}{N}$	$\bar{X}$= Arithmetic Mean ∑X = Summation of values of Variable X N = Number of observations
Short-cut Method	$(\bar{X}) = A + \frac{\sum X}{N}$	A = Assumed Mean ∑d = Sum of deviations of variables from assumed mean
Step Deviation Method	$(\bar{X}) = A + \frac{\sum X}{N} \times C$	∑d' = Sum of step deviations C = Common Factor
Discrete Series **Direct Method**	$(\bar{X}) = \frac{\sum fX}{\sum f}$	∑fX = Sum of product of Variable (X) and frequencies (f) ∑f = Total of frequencies

Short-cut Method	$(\bar{X})$ =A + $\frac{\sum fd}{\sum f}$	∑fd = Sum of product of deviations (d) and respective frequencies (f)
Step Deviation Method	$(\bar{X})$ =A + $\frac{\sum fd'}{\sum f} \times C$	∑fd'= Sum of product of step deviations (d') and respective frequencies (f)
Continuous Series **Direct Method**	$(\bar{X})$ = $\frac{\sum fm}{\sum f}$	m = Mid-Points ∑fm = Sum of product of mid-points (m) and frequencies (f)
Short-cut Method	$(\bar{X})$ = A + $\frac{\sum fd}{\sum f}$	∑fd = Sum of product of deviations (d) from mid-points with the respective frequencies (f)
Step Deviation Method	$(\bar{X})$ = A + $\frac{\sum fd'}{\sum f} \times C$	∑fd' = Sum of product of step deviations (d') and frequencies (f)
2. Combined Mean	$(\bar{X}_{1,2})$ = $\frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1+N_2}$	$\bar{X}_{1,2}$= Combined Mean $\bar{X}_1$ = Arithmetic Mean of first distribution X_2 = Arithmetic Mean of second distribution N_1 = Number of items of first distribution N_2 = Number of items of second distribution
3. Weighted Mean	$\bar{X}_w$ = $\frac{\sum WX}{\sum W}$	X_w = Weighted Mean ∑WX = Sum of product of items and respective weights ∑W = Sum of the weights

MEASURES OF CENTRAL TENDENCY — MEDIAN AND MODE

1. MEDIAN **Individual Series**	Me = Size of $\left[\frac{N+1}{2}\right]^{th}$ item (in case of Odd Number Series)
	Average of two items lying on either side of $\left[\frac{N+1}{2}\right]^{th}$ (in case of Even Number Series)

Discrete Series	Me = Size of $\left[\frac{N+1}{2}\right]^{th}$ item
Continuous Series	Determine Median Class as $\left[\frac{N}{2}\right]^{th}$ item and apply the formula:
	Me = $l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i$

2. LOWER QUARTILE **Individual Series**	Q_1 = Size of $\left[\frac{N+1}{4}\right]^{th}$ item			
Discrete Series	Q_1 = Size of $\left[\frac{N+1}{4}\right]^{th}$ item			
Continuous Series	Determine Quartile Class as $\left[\frac{N}{4}\right]^{th}$ item and apply the formula:			
	$Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f}$			

3. UPPER QUARTILE **Individual Series**	Q_3 = Size of 3 $\left[\frac{N+1}{4}\right]^{th}$ item
Discrete Series	Q_3 = Size of 3 $\left[\frac{N+1}{4}\right]^{th}$ item
Continuous Series	Determine Quartile Class as 3 $\left[\frac{N}{4}\right]^{th}$ item and apply the formula:
	$Q_3 = l_1 + \frac{\frac{3N}{4} - c.f.}{f} \times i$
	Me = Median
	Q_1 = Lower Quartile
	Q_3 = Upper Quartile
	l_1 = Lower limit of the median or Quartile class
	c.f. = Cumulative frequency of the class preceding the median or Quartile class
	f = Simple frequency of the median or Quartile class
	i = Class-intervals

	N = Number of items
4. MODE **Individual Series**	Mode is the value, which occurs largest number of times.
Discrete Series	If the frequencies are regular and homogeneous and there is a single maximum frequency, then Mode is the value corresponding to the highest frequency (Otherwise use Grouping Method)
Continuous Series	Step 1: Determine the Modal Class: (i) By inspection, if frequencies are regular, homogeneous and there is a single maximum frequency; Otherwise (ii) Grouping Method.
	Step 2: Apply the following formula: Mo = $l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$
	Mo = Mode l_1 = Lower limit of the modal class f_1 = Frequency of the modal class f_0 = Frequency of the class preceding the modal class f_2 = Frequency of the class succeeding the modal class i = Class-interval of the modal class

MEASURES OF DISPERSION

1. RANGE	Range = L - S			

Absolute **Measure**	
Relative **Measure**	Coefficient of Range = $\frac{L-S}{L+S}$
	Where, L = Largest item, and S = Smallest item
2. QUARTILE DEVIATION	Interquartile Range = $Q_3 - Q_1$
Absolute **Measure**	Quartile Deviation = $\frac{Q_3 - Q_1}{2}$
Relative **Measure**	Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$
	Where, Q_1 = Lower Quartile; Q_3 = Upper Quartile

3. MEAN DEVIATION		Individual Series		Discrete Series	Continuous Series
Absolute **Measure**	Mean Deviation from Mean $(MD_{\bar{X}})$	$\frac{\sum\lvert x - \bar{X}\rvert}{N}$	$= \frac{\sum\lvert D\rvert}{N}$	$\frac{\sum f\lvert X-\bar{X}\rvert}{N} = \frac{\sum f\lvert D\rvert}{N}$	$\frac{\sum f\lvert m-\bar{X}\rvert}{N} = \frac{\sum f\lvert D\rvert}{N}$

	Mean Deviation from Median (MD$_{Me}$)	$\frac{\sum\lvert x-Me\rvert}{N}$	$=\frac{\sum\lvert D\rvert}{N}$	$\frac{\sum f\lvert X-\mathbf{Me}\rvert}{\mathbf{N}}=\frac{\sum f\lvert D\rvert}{N}$	$\frac{\sum f\lvert m-\mathbf{Me}\rvert}{\mathbf{N}}=\frac{\sum f\lvert D\rvert}{N}$		
					Where, m = Mid-point		
Relative **Measure**	Coefficient of Mean Deviation from Mean ($\bar{X}$) = $\frac{MD_{\bar{X}}}{\bar{X}}$						
	Coefficient of Mean Deviation from Median (Me) = $\frac{MD_{Me}}{Me}$						
4. STANDARD **Deviation**	(σ)	Individual Series		Discrete Series		Continuous Series	
Absolute **Measure**	Actual Mean Method	σ =	$\sqrt{\frac{\sum X^2}{N}}$	σ =	$\sqrt{\frac{\sum fx^2}{N}}$	σ =	$\sqrt{\frac{\sum fx^2}{N}}$
	Direct Method	σ =	$\sqrt{\frac{\sum X_2}{N}-(\bar{\mathbf{X}})^2}$	σ =	$\sqrt{\frac{\sum fx^2}{N}-(\bar{\mathbf{X}})^2}$	σ =	$\sqrt{\frac{\sum fm^2}{N}-(\bar{\mathbf{X}})^2}$
	Short-Cut Method	σ =	$\sqrt{\frac{\sum d^2}{N}-\left(\frac{\sum d}{N}\right)^2}$	σ =	$\sqrt{\frac{\sum fd^2}{N}-\left(\frac{\sum fd}{N}\right)^2}$	σ =	$\sqrt{\frac{\sum fd^2}{N}-\left(\frac{\sum fd}{N}\right)^2}$

	Step Deviation Method	σ = $\sqrt{\frac{\sum f d'^2}{N} - \left(\frac{\sum f d'}{N}\right)^2}$	σ = $\sqrt{\frac{\sum f d'^2}{N} - \left(\frac{\sum f d'}{N}\right)^2} \times C$
	Variance = σ_2		
Relative Measure	Coefficient of Standard Deviation = $\frac{\sigma}{\overline{X}}$		
	Coefficient of Variation (C.V.) = $\frac{\sigma}{\overline{X}}$ × 100		
5. COMBINED STANDARD DEVIATION	Two Related Groups: $\sigma_{1,2}$ =	$\sqrt{\frac{N_1\sigma_1^2\ +\ N_2\sigma_2^2\ +\ N_2d_2^2}{N_1\ +\ N_2}}$	
	$d_1 = \overline{X}_1 - \overline{X}_{1,2}$		
	$d_2 = \overline{X}_1 - \overline{X}_{1,2}$		
	Three Related Groups: $\sigma_{1,2,3}$ = $\sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1\sigma_1^2 + N_2d_2^2\ N_3\sigma_3^2}{N_1 + N_2 + N_3}}$		
	$d_1 = \overline{X}_1 - \overline{X}_{1,2,3}$ $d_2 = \overline{X}_2 - \overline{X}_{1,2,3}$ $d_3 = \overline{X}_3 - \overline{X}_{1,2,3}$		

MEASURES OF CORRELATION

KARL PEARSON’S COEFFICIENT OF CORRELATION	
1. Actual Mean Method	r = $\frac{\sum xy}{N\times\sigma_x\times\sigma_y}=\frac{\sum xy}{N\times\sqrt{\frac{\sum X^2}{N}}\times\sqrt{\frac{\sum Y^2}{N}}}=\frac{\sum XY}{\sqrt{\sum x^2\ \times\sum y^2}}$

	R = Karl Pearson’s Coefficient of Correlation
	N = Number of pair of observations
	X = Deviation of X series form (X - $\bar{X}$)
	V = Deviation of Y series from Mean (Y - $\bar{Y}$)

	σ_x = Standard deviation of X series, i.e., $\sqrt{\frac{\sum X^2}{N}}$
	σ_y = Standard deviation of Y series, i.e., $\sqrt{\frac{\sum Y^2}{N}}$
2. Direct Method	r = $\frac{N\sum XY-\sum X.\sum Y}{\sqrt{N\sum X^2-(\sum X)^2}\times\sqrt{N\sum y^2-(\sum Y)^2}}$
3. Short - Cut Method	r = $\frac{N\sum dxdy-\sum dx\times\sum dy}{\sqrt{N\sum dx^2-(\sum dx)^2}\times\sqrt{N\sum dy^2-(\sum dy)^2}}$

	$\sum dx$ = Sum of deviations of X values from assumed mean. $\sum dy$ = Sum of deviations of Y values from assumed mean. $\sum dx^2$ = Sum of squared deviations of X values from assumed mean. $\sum dy^2$ = Sum of squared deviations of Y values from assumed mean. $\sum dx\ dy$ = Sum of the products of deviations dx and dy.
4. Step Deviation Method	r = $\frac{N\sum dx'dy - \sum dx' \times \sum dy'}{\sqrt{N\sum dx'^2 - (\sum dx')^2} \times \sqrt{N\sum dy'^2 - (\sum dy')^2}}$
	$\sum dx'$ = Sum of step deviations of X values from assumed mean . $\sum dy'$ = Sum of step deviations of Y values from assumed mean. $\sum dx'^2$ = Sum of squared step deviations of X values from assumed mean. $\sum dy'^2$ = Sum of squared step deviations of Y values from assumed mean. $\sum dx'dy'$ = Sum of the products of step deviations dx' and dy'.

Spearman’s Rank Correlation Coefficient	
When Ranks are not Equal	$r_k = 1 - \frac{6\sum D^2}{N^3 - N}$
When Ranks are Equal	$r_k = 1 - \frac{6\left(\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\right)}{N^3 - N}$
	r_k = Coefficient of rank correlation. $\sum D^2$ = Sum of square of rank differences, m = Number of times an item is assigned equal rank.

INDEX NUMBERS

1. UNWEIGHTED INDEX NUMBERS **Simple Aggregative Method**	$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$
Simple Average of Price Relatives	$P_{01} \frac{\Sigma\left(\frac{P_1}{P_0}\times 100\right)}{N}$
2. WEIGHTED INDEX NUMBERS **Weighted Aggregative Method**	
Laspeyre's Method	$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$
Paasche's Method	$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$
Fisher’s Method	$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$
Weighted Average of Price Relatives Method	$P_{01} = \frac{\Sigma RW}{\Sigma W}$

3. CONSUMER PRICE INDEX (CPI) **Aggregate Expenditure Method**	CPI = $\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$
Family Budget Method	CPI = $\frac{\Sigma RW}{\Sigma W}$
4. INDUSTRIAL PRODUCTION INDEX (IPI)	IPI = $\frac{\Sigma\left(\frac{P_1}{P_0} \times 100\right)}{N}$