CONTENTS **Page**

1. DEFINITIONS, SCOPE AND LIMITATIONS

Introduction:

In the modern world of computers and information technology, the importance of statistics is very well recogonised by all the disciplines. Statistics has originated as a science of statehood and found applications slowly and steadily in Agriculture, Economics, Commerce, Biology, Medicine, Industry, planning, education and so on. As on date there is no other human walk of life, where statistics cannot be applied.

Origin and Growth of Statistics:

The word ‘ Statistics’ and ‘ Statistical’ are all derived from the Latin word **Status,** means **a political state**. The theory of statistics as a distinct branch of scientific method is of comparatively recent growth. Research particularly into the mathematical theory of statistics is rapidly proceeding and fresh discoveries are being made all over the world.

Meaning of Statistics:

Statistics is concerned with scientific methods for collecting, organising, summarising, presenting and analysing data as well as deriving valid conclusions and making reasonable decisions on the basis of this analysis. Statistics is concerned with the systematic collection of numerical data and its interpretation. The word ‘ statistic’ is used to referto

1. Numerical facts, such as the number of people living in particular area.
2. The study of ways of collecting, analysing and interpreting the facts.

Definitions:

Statistics is defined differently by different authors over a period of time. In the olden days statistics was confined to only state affairs but in modern days it embraces almost every sphere of

human activity. Therefore a number of old definitions, which was confined to narrow field of enquiry were replaced by more definitions, which are much more comprehensive and exhaustive. Secondly, statistics has been defined in two different ways – Statistical data and statistical methods. The following are some of the definitions of statistics as numerical data.

1. Statistics are the classified facts representing the conditions of people in a state. In particular they are the facts, which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement.
2. Statistics are measurements, enumerations or estimates of natural phenomenon usually systematically arranged, analysed and presented as to exhibit important inter-relationships among them.

Definitions by A.L. Bowley:

Statistics are numerical statement of facts in any department of enquiry placed in relation to each other. - A.L. Bowley

Statistics may be called the science of counting in one of the departments due to Bowley, obviously this is an incomplete definition as it takes into account only the aspect of collection and ignores other aspects such as analysis, presentation and interpretation.

Bowley gives another definition for statistics, which states ‘ statistics may be rightly called the scheme of averages’ . This definition is also incomplete, as averages play an important role in understanding and comparing data and statistics provide more measures.

Definition by Croxton and Cowden:

Statistics may be defined as the science of collection, presentation analysis and interpretation of numerical data from the logical analysis. It is clear that the definition of statistics by Croxton and Cowden is the most scientific and realistic one.

According to this definition there are four stages:

1. Collection of Data: It is the first step and this is the foundation upon which the entire data set. Careful planning is essential before collecting the data. There are different methods of collection of

data such as census, sampling, primary, secondary, etc., and the investigator should make use of correct method.

2. Presentation of data: The mass data collected should be presented in a suitable, concise form for further analysis. The collected data may be presented in the form of tabular or diagrammatic or graphic form.

3. Analysis of data: The data presented should be carefully analysed for making inference from the presented data such as measures of central tendencies, dispersion, correlation, regression etc.,

4. Interpretation of data: The final step is drawing conclusion from the data collected. A valid conclusion must be drawn on the basis of analysis. A high degree of skill and experience is necessary for the interpretation.

Definition by Horace Secrist:

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

The above definition seems to be the most comprehensive and exhaustive.

Functions of Statistics:

There are many functions of statistics. Let us consider the following five important functions.

Condensation:

Generally speaking by the word ‘ to condense’ , we mean to reduce or to lessen. Condensation is mainly applied at embracing the understanding of a huge mass of data by providing only few observations. If in a particular class in Chennai School, only marks in an examination are given, no purpose will be served. Instead if we are given the average mark in that particular examination, definitely it serves the better purpose. Similarly the range of marks is also another measure of the data. Thus, Statistical measures help

to reduce the complexity of the data and consequently to understand any huge mass of data.

Comparison:

Classification and tabulation are the two methods that are used to condense the data. They help us to compare data collected from different sources. Grand totals, measures of central tendency measures of dispersion, graphs and diagrams, coefficient of correlation etc provide ample scope for comparison.

As statistics is an aggregate of facts and figures, comparison is always possible and in fact comparison helps us to understand the data in a better way.

Forecasting:

By the word forecasting, we mean to predict or to estimate before hand. Given the data of the last ten years connected to rainfall of a particular district in Tamilnadu, it is possible to predict or forecast the rainfall for the near future. In business also forecasting plays a dominant role in connection with production, sales, profits etc. The analysis of time series and regression analysis plays an important role in forecasting.

Estimation:

One of the main objectives of statistics is drawn inference about a population from the analysis for the sample drawn from that population. The four major branches of statistical inference are

1. Estimation theory
2. Tests of Hypothesis
3. Non Parametric tests
4. Sequential analysis

In estimation theory, we estimate the unknown value of the population parameter based on the sample observations. Suppose we are given a sample of heights of hundred students in a school, based upon the heights of these 100 students, it is possible to estimate the average height of all students in that school.

Tests of Hypothesis:

A statistical hypothesis is some statement about the probability distribution, characterising a population on the basis of the information available from the sample observations. In the formulation and testing of hypothesis, statistical methods are extremely useful. Whether crop yield has increased because of the use of new fertilizer or whether the new medicine is effective in eliminating a particular disease are some examples of statements of hypothesis and these are tested by proper statistical tools.

Scope of Statistics:

Statistics is not a mere device for collecting numerical data, but as a means of developing sound techniques for their handling, analysing and drawing valid inferences from them. Statistics is applied in every sphere of human activity – social as well as physical – like Biology, Commerce, Education, Planning, Business Management, Information Technology, etc. It is almost impossible to find a single department of human activity where statistics cannot be applied. We now discuss briefly the applications of statistics in other disciplines.

Statistics and Industry:

Statistics is widely used in many industries. In industries, control charts are widely used to maintain a certain quality level. In production engineering, to find whether the product is conforming to specifications or not, statistical tools, namely inspection plans, control charts, etc., are of extreme importance. In inspection plans we have to resort to some kind of sampling – a very important aspect of Statistics.

Statistics and Commerce:

Statistics are lifeblood of successful commerce. Any businessman cannot afford to either by under stocking or having overstock of his goods. In the beginning he estimates the demand for his goods and then takes steps to adjust with his output or purchases. Thus statistics is indispensable in business and commerce.

As so many multinational companies have invaded into our Indian economy, the size and volume of business is increasing. On one side the stiff competition is increasing whereas on the other side the tastes are changing and new fashions are emerging. In this

connection, market survey plays an important role to exhibit the present conditions and to forecast the likely changes in future.

Statistics and Agriculture:

Analysis of variance (ANOVA) is one of the statistical tools developed by Professor R.A. Fisher, plays a prominent role in agriculture experiments. In tests of significance based on small samples, it can be shown that statistics is adequate to test the significant difference between two sample means. In analysis of variance, we are concerned with the testing of equality of several population means.

For an example, five fertilizers are applied to five plots each of wheat and the yield of wheat on each of the plots are given. In such a situation, we are interested in finding out whether the effect of these fertilisers on the yield is significantly different or not. In other words, whether the samples are drawn from the same normal population or not. The answer to this problem is provided by the technique of ANOVA and it is used to test the homogeneity of several population means.

Statistics and Economics:

Statistical methods are useful in measuring numerical changes in complex groups and interpreting collective phenomenon. Nowadays the uses of statistics are abundantly made in any economic study. Both in economic theory and practice, statistical methods play an important role.

Alfred Marshall said, " Statistics are the straw only which I like every other economist have to make the bricks". It may also be noted that statistical data and techniques of statistical tools are immensely useful in solving many economic problems such as wages, prices, production, distribution of income and wealth and so on. Statistical tools like Index numbers, time series Analysis, Estimation theory, Testing Statistical Hypothesis are extensively used in economics.

Statistics and Education:

Statistics is widely used in education. Research has become a common feature in all branches of activities. Statistics is necessary for the formulation of policies to start new course, consideration of facilities available for new courses etc. There are

many people engaged in research work to test the past knowledge and evolve new knowledge. These are possible only through statistics.

Statistics and Planning:

Statistics is indispensable in planning. In the modern world, which can be termed as the "world of planning", almost all the organisations in the government are seeking the help of planning for efficient working, for the formulation of policy decisions and execution of the same.

In order to achieve the above goals, the statistical data relating to production, consumption, demand, supply, prices, investments, income expenditure etc and various advanced statistical techniques for processing, analysing and interpreting such complex data are of importance. In India statistics play an important role in planning, commissioning both at the central and state government levels.

Statistics and Medicine:

In Medical sciences, statistical tools are widely used. In order to test the efficiency of a new drug or medicine, t - test is used or to compare the efficiency of two drugs or two medicines, t-test for the two samples is used. More and more applications of statistics are at present used in clinical investigation.

Statistics and Modern applications:

Recent developments in the fields of computer technology and information technology have enabled statistics to integrate their models and thus make statistics a part of decision making procedures of many organisations. There are so many software packages available for solving design of experiments, forecasting simulation problems etc.

SYSTAT, a software package offers mere scientific and technical graphing options than any other desktop statistics package. SYSTAT supports all types of scientific and technical research in various diversified fields as follows

1. Archeology: Evolution of skull dimensions
2. Epidemiology: Tuberculosis
3. Statistics: Theoretical distributions
4. Manufacturing: Quality improvement

5. Medical research: Clinical investigations.
6. Geology: Estimation of Uranium reserves from ground water

Limitations of statistics:

Statistics with all its wide application in every sphere of human activity has its own limitations. Some of them are given below.

1. **Statistics is not suitable to the study of qualitative phenomenon:** Since statistics is basically a science and deals with a set of numerical data, it is applicable to the study of only these subjects of enquiry, which can be expressed in terms of quantitative measurements. As a matter of fact, qualitative phenomenon like honesty, poverty, beauty, intelligence etc, cannot be expressed numerically and any statistical analysis cannot be directly applied on these qualitative phenomenons. Nevertheless, statistical techniques may be applied indirectly by first reducing the qualitative expressions to accurate quantitative terms. For example, the intelligence of a group of students can be studied on the basis of their marks in a particular examination.
2. **Statistics does not study individuals:** Statistics does not give any specific importance to the individual items, in fact it deals with an aggregate of objects. Individual items, when they are taken individually do not constitute any statistical data and do not serve any purpose for any statistical enquiry.
3. **Statistical laws are not exact:** It is well known that mathematical and physical sciences are exact. But statistical laws are not exact and statistical laws are only approximations. Statistical conclusions are not universally true. They are true only on an average.
4. **Statistics table may be misused:** Statistics must be used only by experts; otherwise, statistical methods are the most dangerous tools on the hands of the inexpert. The use of statistical tools by the inexperienced and untraced persons might lead to wrong conclusions. Statistics can be easily misused by quoting wrong figures of data. As King says

aptly ‘ statistics are like clay of which one can make a God or Devil as one pleases’ .

5. **Statistics is only, one of the methods of studying a problem:**

Statistical method do not provide complete solution of the problems because problems are to be studied taking the background of the countries culture, philosophy or religion into consideration. Thus the statistical study should be supplemented by other evidences.

Exercise – 1

I. Choose the best answer:

1. The origin of statistics can be traced to
   (a) State (b) Commerce
   (c) Economics (d) Industry.
2. ‘ Statistics may be called the science of counting’ is the definition given by
   (a) Croxton (b) A.L.Bowley
   (c) Boddington (d) Webster.

II. Fill in the blanks:

3. In the olden days statistics was confined to only_____.
4. Classification and_____are the two methods that are used to condense the data.
5. The analysis of time series and regression analysis plays an important role in_____.
6. _____is one of the statistical tool plays prominent role in agricultural experiments.

Answers:

I. 1. (a)

2. (b)

II. 3. State affairs

4. Tabulation

5. Forecasting

6. Analysis of variance (or ANOVA)

2. INTRODUCTION TO SAMPLING METHODS

Introduction:

Sampling is very often used in our daily life. For example while purchasing food grains from a shop we usually examine a handful from the bag to assess the quality of the commodity. A doctor examines a few drops of blood as sample and draws conclusion about the blood constitution of the whole body. Thus most of our investigations are based on samples. In this chapter, let us see the importance of sampling and the various methods of sample selections from the population.

Population:

In a statistical enquiry, all the items, which fall within the purview of enquiry, are known as **Population** or **Universe**. In other words, the population is a complete set of all possible observations of the type which is to be investigated. Total number of students studying in a school or college, total number of books in a library, total number of houses in a village or town are some examples of population.

Sometimes it is possible and practical to examine every person or item in the population we wish to describe. We call this a **Complete enumeration,** or **census**. We use **sampling** when it is not possible to measure every item in the population. Statisticians use the word population to refer not only to people but to all items that have been chosen for study.

Finite population and infinite population:

A population is said to be finite if it consists of finite number of units. Number of workers in a factory, production of articles in a particular day for a company are examples of finite population. The total number of units in a population is called population size. A population is said to be infinite if it has infinite number of units. For example the number of stars in the sky, the number of people seeing the Television programmes etc.,

Census Method:

Information on population can be collected in two ways – census method and sample method. In census method every element of the population is included in the investigation. For example, if we study the average annual income of the families of a particular village or area, and if there are 1000 families in that area, we must study the income of all 1000 families. In this method no family is left out, as each family is a unit.

Population census of India:

The population census of our country is taken at 10 yearly intervals. The latest census was taken in 2001. The first census was taken in 1871 – 72.

Merits and limitations of Census method:

Mertis:

1. The data are collected from each and every item of the population
2. The results are more accurate and reliable, because every item of the universe is required.
3. Intensive study is possible.
4. The data collected may be used for various surveys, analyses etc.

Limitations:

1. It requires a large number of enumerators and it is a costly method
2. It requires more money, labour, time energy etc.
3. It is not possible in some circumstances where the universe is infinite.

Sampling:

The theory of sampling has been developed recently but this is not new. In our everyday life we have been using sampling theory as we have discussed in introduction. In all those cases we believe that the samples give a correct idea about the population. Most of our decisions are based on the examination of a few items that is sample studies.

Sample:

Statisticians use the word **sample** to describe a portion chosen from the population. A finite subset of statistical individuals defined in a population is called a sample. The number of units in a sample is called the **sample size.**

Sampling unit:

The constituents of a population which are individuals to be sampled from the population and cannot be further subdivided for the purpose of the sampling at a time are called sampling units. For example, to know the average income per family, the head of the family is a sampling unit. To know the average yield of rice, each farm owner' s yield of rice is a sampling unit.

Sampling frame:

For adopting any sampling procedure it is essential to have a list identifying each sampling unit by a number. Such a list or map is called sampling frame. A list of voters, a list of house holders, a list of villages in a district, a list of farmers etc. are a few examples of sampling frame.

Reasons for selecting a sample:

Sampling is inevitable in the following situations:

1. Complete enumerations are practically impossible when the population is infinite.
2. When the results are required in a short time.
3. When the area of survey is wide.
4. When resources for survey are limited particularly in respect of money and trained persons.
5. When the item or unit is destroyed under investigation.

Parameters and statistics:

We can describe samples and populations by using measures such as the mean, median, mode and standard deviation. When these terms describe the characteristics of a population, they are called **parameters**. When they describe the characteristics of a sample, they are called **statistics**. A parameter is a characteristic of a population and a statistic is a characteristic of a sample. Since samples are subsets of population statistics provide estimates of the

parameters. That is, when the parameters are unknown, they are estimated from the values of the statistics.

In general, we use Greek or capital letters for population parameters and lower case Roman letters to denote sample statistics. [N, μ, σ, are the standard symbols for the size, mean, S.D, of population. n , $\bar{x}$, s, are the standard symbol for the size, mean, s.d of sample respectively].

Principles of Sampling:

Samples have to provide good estimates. The following principle tell us that the sample methods provide such good estimates

1. Principle of statistical regularity:

A moderately large number of units chosen at random from a large group are almost sure on the average to possess the characteristics of the large group.

2. Principle of Inertia of large numbers:

Other things being equal, as the sample size increases, the results tend to be more accurate and reliable.

3. Principle of Validity:

This states that the sampling methods provide valid estimates about the population units (parameters).

4. Principle of Optimisation:

This principle takes into account the desirability of obtaining a sampling design which gives optimum results. This minimizes the risk or loss of the sampling design.

The foremost purpose of sampling is to gather maximum information about the population under consideration at minimum cost, time and human power. This is best achieved when the sample contains all the properties of the population.

Sampling errors and non-sampling errors:

The two types of errors in a sample survey are sampling errors and non - sampling errors.

1. Sampling errors:

Although a sample is a part of population, it cannot be expected generally to supply full information about population. So there may be in most cases difference between statistics and

parameters. The discrepancy between a parameter and its estimate due to sampling process is known as **sampling error.**

2. Non-sampling errors:

In all surveys some errors may occur during collection of actual information. These errors are called Non-sampling errors**.**

Advantages and Limitation of Sampling:

There are many advantages of sampling methods over census method. They are as follows:

1. Sampling saves time and labour.
2. It results in reduction of cost in terms of money and man-hour.
3. Sampling ends up with greater accuracy of results.
4. It has greater scope.
5. It has greater adaptability.
6. If the population is too large, or hypothetical or destroyable sampling is the only method to be used.

The limitations of sampling are given below:

1. Sampling is to be done by qualified and experienced persons. Otherwise, the information will be unbelievable.
2. Sample method may give the extreme values sometimes instead of the mixed values.
3. There is the possibility of sampling errors. Census survey is free from sampling error.

Types of Sampling:

The technique of selecting a sample is of fundamental importance in sampling theory and it depends upon the nature of investigation. The sampling procedures which are commonly used may be classified as

1. Probability sampling.
2. Non-probability sampling.
3. Mixed sampling.

Probability sampling (Random sampling):

A probability sample is one where the selection of units from the population is made according to known probabilities. (eg.) Simple random sample, probability proportional to sample size etc.

Non-Probability sampling:

It is the one where discretion is used to select ‘ representative’ units from the population (or) to infer that a sample is ‘ representative’ of the population. This method is called **judgement or purposive** sampling. This method is mainly used for opinion surveys; A common type of judgement sample used in surveys is quota sample. This method is not used in general because of prejudice and bias of the enumerator. However if the enumerator is experienced and expert, this method may yield valuable results. For example, in the market research survey of the performance of their new car, the sample was all new car purchasers.

Mixed Sampling:

Here samples are selected partly according to some probability and partly according to a fixed sampling rule; they are termed as mixed samples and the technique of selecting such samples is known as **mixed sampling**.

Methods of selection of samples:

Here we shall consider the following three methods:

1. Simple random sampling.
2. Stratified random sampling.
3. Systematic random sampling.

1. Simple random sampling:

A simple random sample from finite population is a sample selected such that each possible sample combination has equal probability of being chosen. It is also called unrestricted random sampling.

2. Simple random sampling without replacement:

In this method the population elements can enter the sample only once (ie) the units once selected is not returned to the population before the next draw.

3. Simple random sampling with replacement:

In this method the population units may enter the sample more than once. Simple random sampling may be with or without replacement.

Methods of selection of a simple random sampling:

The following are some methods of selection of a simple random sampling.

a) Lottery Method:

This is the most popular and simplest method. In this method all the items of the population are numbered on separate slips of paper of same size, shape and colour. They are folded and mixed up in a container. The required numbers of slips are selected at random for the desire sample size. For example, if we want to select 5 students, out of 50 students, then we must write their names or their roll numbers of all the 50 students on slips and mix them. Then we make a random selection of 5 students.

This method is mostly used in lottery draws. If the universe is infinite this method is inapplicable.

b) Table of Random numbers:

As the lottery method cannot be used, when the population is infinite, the alternative method is that of using the table of random numbers. There are several standard tables of random numbers.

1. Tippett' s table
2. Fisher and Yates' table
3. Kendall and Smith' s table are the three tables among them.

A random number table is so constructed that all digits 0 to 9 appear independent of each other with equal frequency. If we have to select a sample from population of size N= 100, then the numbers can be combined three by three to give the numbers from 001 to 100.

Procedure to select a sample using random number table:

Units of the population from which a sample is required are assigned with equal number of digits. When the size of the population is less than thousand, three digit number 000,001,002, …. 999 are assigned. We may start at any place and may go on in any direction such as column wise or row- wise in a random number table. But consecutive numbers are to be used.

On the basis of the size of the population and the random number table available with us, we proceed according to our

convenience. If any random number is greater than the population size N, then N can be subtracted from the random number drawn. This can be repeatedly until the number is less than N or equal to N.

Example 1:

In an area there are 500 families.Using the following extract from a table of random numbers select a sample of 15 families to find out the standard of living of those families in that area.

4652	3819	8431	2150	2352	2472	0043	3488
9031	7617	1220	4129	7148	1943	4890	1749
2030	2327	7353	6007	9410	9179	2722	8445
0641	1489	0828	0385	8488	0422	7209	4950

Solution:

In the above random number table we can start from any row or column and read three digit numbers continuously row-wise or column wise.

Now we start from the third row, the numbers are:

203 023 277 353 600 794 109 179

272 284 450 641 148 908 280

Since some numbers are greater than 500, we subtract 500 from those numbers and we rewrite the selected numbers as follows:

203 023 277 353 100 294 109 179

272 284 450 141 148 408 280

c) Random number selections using calculators or computers:

Random number can be generated through scientific calculator or computers. For each press of the key get a new random numbers. The ways of selection of sample is similar to that of using random number table.

Merits of using random numbers:

Merits:

1. Personal bias is eliminated as a selection depends solely on chance .
2. A random sample is in general a representative sample for a homogenous population.
3. There is no need for the thorough knowledge of the units of the population.
4. The accuracy of a sample can be tested by examining another sample from the same universe when the universe is unknown.
5. This method is also used in other methods of sampling.

Limitations:

1. Preparing lots or using random number tables is tedious when the population is large.
2. When there is large difference between the units of population, the simple random sampling may not be a representative sample.
3. The size of the sample required under this method is more than that required by stratified random sampling.
4. It is generally seen that the units of a simple random sample lie apart geographically. The cost and time of collection of data are more.

Stratified Random Sampling:

Of all the methods of sampling the procedure commonly used in surveys is stratified sampling. This technique is mainly used to reduce the population heterogeneity and to increase the efficiency of the estimates. Stratification means division into groups.In this method the population is divided into a number of subgroups or strata. The strata should be so formed that each stratum is homogeneous as far as possible. Then from each stratum a simple random sample may be selected and these are combined together to form the required sample from the population.

Types of Stratified Sampling:

There are two types of stratified sampling. They are **proportional** and **non-proportional**. In the proportional sampling

equal and proportionate representation is given to subgroups or strata. If the number of items is large, the sample will have a higher size and vice versa.

The population size is denoted by N and the sample size is denoted by ' n' the sample size is allocated to each stratum in such a way that the sample fractions is a constant for each stratum. That is given by n/N = c. So in this method each stratum is represented according to its size.

In non-proportionate sample, equal representation is given to all the sub-strata regardless of their existence in the population.

Example 2:

A sample of 50 students is to be drawn from a population consisting of 500 students belonging to two institutions A and B. The number of students in the institution A is 200 and the institution B is 300. How will you draw the sample using proportional allocation?

Solution:

There are two strata in this case with sizes $N_1 = 200$ and $N_2 = 300$ and the total population $N = N_1 + N_2 = 500$

The sample size is 50.

If n_1 and n_2 are the sample sizes,

$$n_1 = \frac{n}{N} \times N_1 = \frac{50}{500} \times 200 = 20$$

$$n_2 = \frac{n}{N} \times N_2 = \frac{50}{500} \times 300 = 30$$

The sample sizes are 20 from A and 30 from B. Then the units from each institution are to be selected by simple random sampling.

Merits and limitations of stratified sampling:

Merits:

1. It is more representative.
2. It ensures greater accuracy.

3. It is easy to administer as the universe is sub - divided.
4. Greater geographical concentration reduces time and expenses.
5. When the original population is badly skewed, this method is appropriate.
6. For non – homogeneous population, it may field good results.

Limitations:

1. To divide the population into homogeneous strata, it requires more money, time and statistical experience which is a difficult one.
2. Improper stratification leads to bias, if the different strata overlap such a sample will not be a representative one.

Systematic Sampling:

This method is widely employed because of its ease and convenience. A frequently used method of sampling when a complete list of the population is available is **systematic sampling**. It is also called **Quasi-random sampling**.

Selection procedure:

The whole sample selection is based on just a random start . The first unit is selected with the help of random numbers and the rest get selected automatically according to some pre designed pattern is known as **systematic sampling**. With systematic random sampling every K^{th} element in the frame is selected for the sample, with the starting point among the first K elements determined at random.

For example, if we want to select a sample of 50 students from 500 students under this method K^{th} item is picked up from the sampling frame and K is called the **sampling interval**.

$$\text{Sampling interval , } K = \frac{N}{n} = \frac{\text{Population size}}{\text{Samplesize}}$$

$$K = \frac{500}{50} = 10$$

K = 10 is the sampling interval. Systematic sample consists in selecting a random number say i K and every K^{th} unit

subsequently. Suppose the random number ' i' is 5, then we select 5, 15, 25, 35, 45,……. The random number ' i' is called random start. The technique will generate K systematic samples with equal probability.

Merits :

1. This method is simple and convenient.
2. Time and work is reduced much.
3. If proper care is taken result will be accurate.
4. It can be used in infinite population.

Limitations:

1. Systematic sampling may not represent the whole population.
2. There is a chance of personal bias of the investigators.

Systematic sampling is preferably used when the information is to be collected from trees in a forest, house in blocks, entries in a register which are in a serial order etc.

Exercise – 2

I. Choose the best Answer:

1. Sampling is inevitable in the situations
   (a) Blood test of a person
   (b) When the population is infinite
   (c) Testing of life of dry battery cells
   (d) All the above
2. The difference between sample estimate and population parameter is termed as
   (a) Human error (b) Sampling error
   (c) Non-sampling error (d) None of the above
3. If each and every unit of population has equal chance of being included in the sample, it is known as
   (a) Restricted sampling (b) Purposive sampling
   (c) Simple random sampling (d) None of the above
4. Simple random sample can be drawn with the help of
   (a) Slip method
   (b) Random number table
   (c) Calculator
   (d) All the above

5. A selection procedure of a sample having no involvement of probability is known as

(a) Purposive sampling (b) Judgement sampling

(c) Subjective sampling (d) All the above

6. Five establishments are to be selected from a list of 50 establishments by systematic random sampling. If the first number is 7, the next one is

(a) 8 (b) 16 (c) 17 (d) 21

II. Fill in the blanks:

7. A population consisting of an unlimited number of units is called an_____population

8. If all the units of a population are surveyed it is called _____

9. The discrepancy between a parameter and its estimate due to sampling process is known as _____

10. The list of all the items of a population is known as _____

11. Stratified sampling is appropriate when population is _____

12. When the items are perishable under investigation it is not possible to do _____

13. When the population consists of units arranged in a sequence would prefer_____sampling

14. For a homogeneous population,_____sampling is better than stratified random sampling.

Answers:

I.

1. (d) 2.(b) 3. (c) 4.(d) 5.(d) 6.(c)

II.

7. infinite
8. complete enumeration or census
9. sampling error
10. sampling frame
11. heterogeneous or Non- homogeneous
12. complete enumeration
13. systematic
14. simple random

3. COLLECTION OF DATA, CLASSIFICATION AND TABULATION

Introduction:

Everybody collects, interprets and uses information, much of it in a numerical or statistical forms in day-to-day life. It is a common practice that people receive large quantities of information everyday through conversations, televisions, computers, the radios, newspapers, posters, notices and instructions. It is just because there is so much information available that people need to be able to absorb, select and reject it. In everyday life, in business and industry, certain statistical information is necessary and it is independent to know where to find it how to collect it. As consequences, everybody has to compare prices and quality before making any decision about what goods to buy. As employees of any firm, people want to compare their salaries and working conditions, promotion opportunities and so on. In time the firms on their part want to control costs and expand their profits.

One of the main functions of statistics is to provide information which will help on making decisions. Statistics provides the type of information by providing a description of the present, a profile of the past and an estimate of the future. The following are some of the objectives of collecting statistical information.

1. To describe the methods of collecting primary statistical information.
2. To consider the status involved in carrying out a survey.
3. To analyse the process involved in observation and interpreting.
4. To define and describe sampling.
5. To analyse the basis of sampling.
6. To describe a variety of sampling methods.

Statistical investigation is a comprehensive and requires systematic collection of data about some group of people or objects, describing and organizing the data, analyzing the data with

the help of different statistical method, summarizing the analysis and using these results for making judgements, decisions and predictions. The validity and accuracy of final judgement is most crucial and depends heavily on how well the data was collected in the first place. The quality of data will greatly affect the conditions and hence at most importance must be given to this process and every possible precautions should be taken to ensure accuracy while collecting the data.

Nature of data:

It may be noted that different types of data can be collected for different purposes. The data can be collected in connection with time or geographical location or in connection with time and location. The following are the three types of data:

1. Time series data.
2. Spatial data
3. Spacio-temporal data.

Time series data:

It is a collection of a set of numerical values, collected over a period of time. The data might have been collected either at regular intervals of time or irregular intervals of time.

Example 1:

The following is the data for the three types of expenditures in rupees for a family for the four years 2001,2002,2003,2004.

Year	Food	Education	Others	Total
2001	3000	2000	3000	8000
2002	3500	3000	4000	10500
2003	4000	3500	5000	12500
2004	5000	5000	6000	16000

Spatial Data:

If the data collected is connected with that of a place, then it is termed as spatial data. For example, the data may be

1. Number of runs scored by a batsman in different test matches in a test series at different places
2. District wise rainfall in Tamilnadu
3. Prices of silver in four metropolitan cities

Example 2:

The population of the southern states of India in 1991.

State	Population
Tamilnadu	5,56,38,318
Andhra Pradesh	6,63,04,854
Karnataka	4,48,17,398
Kerala	2,90,11,237
Pondicherry	7,89,416

Spacio Temporal Data:

If the data collected is connected to the time as well as place then it is known as spacio temporal data.

Example 3:

State	Population	
	1981	1991
Tamil Nadu	4,82,97,456	5,56,38,318
Andhra Pradesh	5,34,03,619	6,63,04,854
Karnataka	3,70,43,451	4,48,17,398
Kerala	2,54,03,217	2,90,11,237
Pondicherry	6,04,136	7,89,416

Categories of data:

Any statistical data can be classified under two categories depending upon the sources utilized.

These categories are,

1. Primary data 2. Secondary data

Primary data:

Primary data is the one, which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by survey conducted by individuals or research institution or any organisation.

Example 4:

If a researcher is interested to know the impact of noon-meal scheme for the school children, he has to undertake a survey and collect data on the opinion of parents and children by asking relevant questions. Such a data collected for the purpose is called primary data.

The primary data can be collected by the following five methods.

1. Direct personal interviews.
2. Indirect Oral interviews.
3. Information from correspondents.
4. Mailed questionnaire method.
5. Schedules sent through enumerators.

1. Direct personal interviews:

The persons from whom informations are collected are known as informants. The investigator personally meets them and asks questions to gather the necessary informations. It is the suitable method for intensive rather than extensive field surveys. It suits best for intensive study of the limited field.

Merits:

1. People willingly supply informations because they are approached personally. Hence, more response noticed in this method than in any other method.
2. The collected informations are likely to be uniform and accurate. The investigator is there to clear the doubts of the informants.
3. Supplementary informations on informant' s personal aspects can be noted. Informations on character and environment may help later to interpret some of the results.
4. Answers for questions about which the informant is likely to be sensitive can be gathered by this method.
5. The wordings in one or more questions can be altered to suit any informant. Explanations may be given in other languages also. Inconvenience and misinterpretations are thereby avoided.

Limitations:

1. It is very costly and time consuming.
2. It is very difficult, when the number of persons to be interviewed is large and the persons are spread over a wide area.
3. Personal prejudice and bias are greater under this method.

2. Indirect Oral Interviews:

Under this method the investigator contacts witnesses or neighbours or friends or some other third parties who are capable of supplying the necessary information. This method is preferred if the required information is on addiction or cause of fire or theft or murder etc., If a fire has broken out a certain place, the persons living in neighbourhood and witnesses are likely to give information on the cause of fire. In some cases, police interrogated third parties who are supposed to have knowledge of a theft or a murder and get some clues. Enquiry committees appointed by governments generally adopt this method and get people' s views and all possible details of facts relating to the enquiry. This method is suitable whenever direct sources do not exists or cannot be relied upon or would be unwilling to part with the information.

The validity of the results depends upon a few factors, such as the nature of the person whose evidence is being recorded, the ability of the interviewer to draw out information from the third parties by means of appropriate questions and cross examinations, and the number of persons interviewed. For the success of this method one person or one group alone should not be relied upon.

3. Information from correspondents:

The investigator appoints local agents or correspondents in different places and compiles the information sent by them. Information to Newspapers and some departments of Government come by this method. The advantage of this method is that it is cheap and appropriate for extensive investigations. But it may not ensure accurate results because the correspondents are likely to be negligent, prejudiced and biased. This method is adopted in those cases where information are to be collected periodically from a wide area for a long time.

4. Mailed questionnaire method:

Under this method a list of questions is prepared and is sent to all the informants by post. The list of questions is technically called questionnaire. A covering letter accompanying the questionnaire explains the purpose of the investigation and the importance of correct information and request the informants to fill in the blank spaces provided and to return the form within a specified time. This method is appropriate in those cases where the informants are literates and are spread over a wide area.

Merits:

1. It is relatively cheap.
2. It is preferable when the informants are spread over the wide area.

Limitations:

1. The greatest limitation is that the informants should be literates who are able to understand and reply the questions.
2. It is possible that some of the persons who receive the questionnaires do not return them.
3. It is difficult to verify the correctness of the informations furnished by the respondents.

With the view of minimizing non-respondents and collecting correct information, the questionnaire should be carefully drafted. There is no hard and fast rule. But the following general principles may be helpful in framing the questionnaire. A covering letter and a self addressed and stamped envelope should accompany the questionnaire. The covering letter should politely point out the purpose of the survey and privilege of the respondent who is one among the few associated with the investigation. It should assure that the information would be kept confidential and would never be misused. It may promise a copy of the findings or free gifts or concessions etc.,

Characteristics of a good questionnaire:

1. Number of questions should be minimum.
2. Questions should be in logical orders, moving from easy to more difficult questions.

3. Questions should be short and simple. Technical terms and vague expressions capable of different interpretations should be avoided.
4. Questions fetching YES or NO answers are preferable. There may be some multiple choice questions requiring lengthy answers are to be avoided.
5. Personal questions and questions which require memory power and calculations should also be avoided.
6. Question should enable cross check. Deliberate or unconscious mistakes can be detected to an extent.
7. Questions should be carefully framed so as to cover the entire scope of the survey.
8. The wording of the questions should be proper without hurting the feelings or arousing resentment.
9. As far as possible confidential informations should not be sought.
10. Physical appearance should be attractive, sufficient space should be provided for answering each questions.

5. Schedules sent through Enumerators:

Under this method enumerators or interviewers take the schedules, meet the informants and filling their replies. Often distinction is made between the schedule and a questionnaire. A schedule is filled by the interviewers in a face-to-face situation with the informant. A questionnaire is filled by the informant which he receives and returns by post. It is suitable for extensive surveys.

Merits:

1. It can be adopted even if the informants are illiterates.
2. Answers for questions of personal and pecuniary nature can be collected.
3. Non-response is minimum as enumerators go personally and contact the informants.
4. The informations collected are reliable. The enumerators can be properly trained for the same.
5. It is most popular methods.

Limitations:

1. It is the costliest method.

2. Extensive training is to be given to the enumerators for collecting correct and uniform informations.
3. Interviewing requires experience. Unskilled investigators are likely to fail in their work.

Before the actual survey, a pilot survey is conducted. The questionnaire/Schedule is pre-tested in a pilot survey. A few among the people from whom actual information is needed are asked to reply. If they misunderstand a question or find it difficult to answer or do not like its wordings etc., it is to be altered. Further it is to be ensured that every questions fetches the desired answer.

Merits and Demerits of primary data:

1. The collection of data by the method of personal survey is possible only if the area covered by the investigator is small. Collection of data by sending the enumerator is bound to be expensive. Care should be taken twice that the enumerator record correct information provided by the informants.
2. Collection of primary data by framing a schedules or distributing and collecting questionnaires by post is less expensive and can be completed in shorter time.
3. Suppose the questions are embarrassing or of complicated nature or the questions probe into personnel affairs of individuals, then the schedules may not be filled with accurate and correct information and hence this method is unsuitable.
4. The information collected for primary data is mere reliable than those collected from the secondary data.

Secondary Data:

Secondary data are those data which have been already collected and analysed by some earlier agency for its own use; and later the same data are used by a different agency. According to W.A.Neiswanger, ‘ A primary source is a publication in which the data are published by the same authority which gathered and analysed them. A secondary source is a publication, reporting the data which have been gathered by other authorities and for which others are responsible’ .

Sources of Secondary data:

In most of the studies the investigator finds it impracticable to collect first-hand information on all related issues and as such he makes use of the data collected by others. There is a vast amount of published information from which statistical studies may be made and fresh statistics are constantly in a state of production. The sources of secondary data can broadly be classified under two heads:

1. Published sources, and
2. Unpublished sources.

1. Published Sources:

The various sources of published data are:

1. Reports and official publications of
   (i) International bodies such as the International Monetary Fund, International Finance Corporation and United Nations Organisation.
   (ii) Central and State Governments such as the Report of the Tandon Committee and Pay Commission.
2. Semi-official publication of various local bodies such as Municipal Corporations and District Boards.
3. Private publications-such as the publications of –
   (i) Trade and professional bodies such as the Federation of Indian Chambers of Commerce and Institute of Chartered Accountants.
   (ii) Financial and economic journals such as ' Commerce' , ' Capital' and ' Indian Finance' .
   (iii) Annual reports of joint stock companies.
   (iv) Publications brought out by research agencies, research scholars, etc.

It should be noted that the publications mentioned above vary with regard to the periodically of publication. Some are published at regular intervals (yearly, monthly, weekly etc.,) whereas others are ad hoc publications, i.e., with no regularity about periodicity of publications.

Note: A lot of secondary data is available in the internet. We can access it at any time for the further studies.

2. Unpublished Sources

All statistical material is not always published. There are various sources of unpublished data such as records maintained by various Government and private offices, studies made by research institutions, scholars, etc. Such sources can also be used where necessary

Precautions in the use of Secondary data

The following are some of the points that are to be considered in the use of secondary data

1. How the data has been collected and processed
2. The accuracy of the data
3. How far the data has been summarized
4. How comparable the data is with other tabulations
5. How to interpret the data, especially when figures collected for one purpose is used for another

Generally speaking, with secondary data, people have to compromise between what they want and what they are able to find.

Merits and Demerits of Secondary Data:

1. Secondary data is cheap to obtain. Many government publications are relatively cheap and libraries stock quantities of secondary data produced by the government, by companies and other organisations.
2. Large quantities of secondary data can be got through internet.
3. Much of the secondary data available has been collected for many years and therefore it can be used to plot trends.
4. Secondary data is of value to:
    - The government – help in making decisions and planning future policy.
    - Business and industry – in areas such as marketing, and sales in order to appreciate the general economic and social conditions and to provide information on competitors.
    - Research organisations – by providing social, economical and industrial information.

Classification:

The collected data, also known as raw data or ungrouped data are always in an un organised form and need to be organised and presented in meaningful and readily comprehensible form in order to facilitate further statistical analysis. It is, therefore, essential for an investigator to condense a mass of data into more and more comprehensible and assimilable form. The process of grouping into different classes or sub classes according to some characteristics is known as classification, tabulation is concerned with the systematic arrangement and presentation of classified data. Thus classification is the first step in tabulation.

For Example, letters in the post office are classified according to their destinations viz., Delhi, Madurai, Bangalore, Mumbai etc.,

Objects of Classification:

The following are main objectives of classifying the data:

1. It condenses the mass of data in an easily assimilable form.
2. It eliminates unnecessary details.
3. It facilitates comparison and highlights the significant aspect of data.
4. It enables one to get a mental picture of the information and helps in drawing inferences.
5. It helps in the statistical treatment of the information collected.

Types of classification:

Statistical data are classified in respect of their characteristics. Broadly there are four basic types of classification namely

a) Chronological classification
b) Geographical classification
c) Qualitative classification
d) Quantitative classification

a) Chronological classification:

In chronological classification the collected data are arranged according to the order of time expressed in years, months, weeks, etc., The data is generally classified in ascending order of

time. For example, the data related with population, sales of a firm, imports and exports of a country are always subjected to chronological classification.

Example 5:

The estimates of birth rates in India during 1970 – 76 are

Year	1970	1971	1972	1973	1974	1975	1976
Birth Rate	36.8	36.9	36.6	34.6	34.5	35.2	34.2

b) Geographical classification:

In this type of classification the data are classified according to geographical region or place. For instance, the production of paddy in different states in India, production of wheat in different countries etc.,

Example 6:

Country	America	China	Denmark	France	India
Yield of wheat in (kg/acre)	1925	893	225	439	862

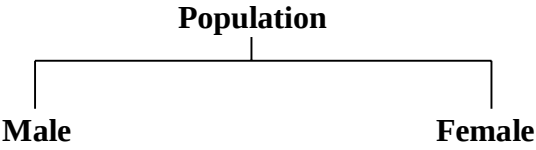
c) Qualitative classification:

In this type of classification data are classified on the basis of same attributes or quality like sex, literacy, religion, employment etc., Such attributes cannot be measured along with a scale.

For example, if the population to be classified in respect to one attribute, say sex, then we can classify them into two namely that of males and females. Similarly, they can also be classified into ‘ employed’ or ‘ unemployed’ on the basis of another attribute ‘ employment’.

Thus when the classification is done with respect to one attribute, which is dichotomous in nature, two classes are formed, one possessing the attribute and the other not possessing the attribute. This type of classification is called simple or dichotomous classification.

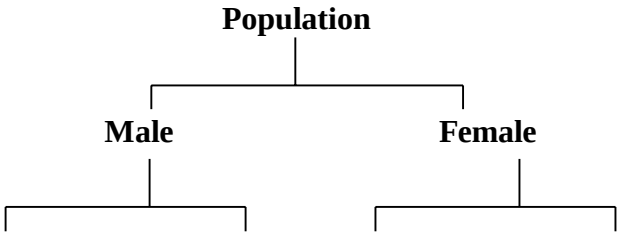
A simple classification may be shown as under



The classification, where two or more attributes are considered and several classes are formed, is called a manifold classification. For example, if we classify population simultaneously with respect to two attributes, e.g sex and employment, then population are first classified with respect to ‘ sex’ into ‘ males’ and ‘ females’ . Each of these classes may then be further classified into ‘ employment’ and ‘ unemployment’ on the basis of attribute ‘ employment’ and as such Population are classified into four classes namely.

(i) Male employed
(ii) Male unemployed
(iii) Female employed
(iv) Female unemployed

Still the classification may be further extended by considering other attributes like marital status etc. This can be explained by the following chart



d) Quantitative classification:

Quantitative classification refers to the classification of data according to some characteristics that can be measured such as height, weight, etc., For example the students of a college may be classified according to weight as given below.

Weight (in lbs)	No of Students
90-100	50
100-110	200
110-120	260
120-130	360
130-140	90
140-150	40
Total	1000

In this type of classification there are two elements, namely (i) the variable (i.e) the weight in the above example, and (ii) the frequency in the number of students in each class. There are 50 students having weights ranging from 90 to 100 lb, 200 students having weight ranging between 100 to 110 lb and so on.

Tabulation:

Tabulation is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information. A table is a systematic arrangement of classified data in columns and rows. Thus, a statistical table makes it possible for the investigator to present a huge mass of data in a detailed and orderly form. It facilitates comparison and often reveals certain patterns in data which are otherwise not obvious.Classification and ‘ Tabulation’ , as a matter of fact, are not two distinct processes. Actually they go together. Before tabulation data are classified and then displayed under different columns and rows of a table.

Advantages of Tabulation:

Statistical data arranged in a tabular form serve following objectives:

1. It simplifies complex data and the data presented are easily understood.
2. It facilitates comparison of related facts.
3. It facilitates computation of various statistical measures like averages, dispersion, correlation etc.

4. It presents facts in minimum possible space and unnecessary repetitions and explanations are avoided. Moreover, the needed information can be easily located.
5. Tabulated data are good for references and they make it easier to present the information in the form of graphs and diagrams.

Preparing a Table:

The making of a compact table itself an art. This should contain all the information needed within the smallest possible space. What the purpose of tabulation is and how the tabulated information is to be used are the main points to be kept in mind while preparing for a statistical table. An ideal table should consist of the following main parts:

1. Table number
2. Title of the table
3. Captions or column headings
4. Stubs or row designation
5. Body of the table
6. Footnotes
7. Sources of data

Table Number:

A table should be numbered for easy reference and identification. This number, if possible, should be written in the centre at the top of the table. Sometimes it is also written just before the title of the table.

Title:

A good table should have a clearly worded, brief but unambiguous title explaining the nature of data contained in the table. It should also state arrangement of data and the period covered. The title should be placed centrally on the top of a table just below the table number (or just after table number in the same line).

Captions or column Headings:

Captions in a table stands for brief and self explanatory headings of vertical columns. Captions may involve headings and

sub-headings as well. The unit of data contained should also be given for each column. Usually, a relatively less important and shorter classification should be tabulated in the columns.

Stubs or Row Designations:

Stubs stands for brief and self explanatory headings of horizontal rows. Normally, a relatively more important classification is given in rows. Also a variable with a large number of classes is usually represented in rows. For example, rows may stand for score of classes and columns for data related to sex of students. In the process, there will be many rows for scores classes but only two columns for male and female students.

A model structure of a table is given below:

Table Number Title of the Table

Sub Heading	Caption Headings Caption Sub-Headings	Total
Stub Sub- Headings	Body	
Total		

Foot notes:
Sources Note:

Body:

The body of the table contains the numerical information of frequency of observations in the different cells. This arrangement of data is according to the discription of captions and stubs.

Footnotes:

Footnotes are given at the foot of the table for explanation of any fact or information included in the table which needs some explanation. Thus, they are meant for explaining or providing further details about the data, that have not been covered in title, captions and stubs.

Sources of data:

Lastly one should also mention the source of information from which data are taken. This may preferably include the name of the author, volume, page and the year of publication. This should also state whether the data contained in the table is of ' primary or secondary' nature.

Requirements of a Good Table:

A good statistical table is not merely a careless grouping of columns and rows but should be such that it summarizes the total information in an easily accessible form in minimum possible space. Thus while preparing a table, one must have a clear idea of the information to be presented, the facts to be compared and he points to be stressed.

Though, there is no hard and fast rule for forming a table yet a few general point should be kept in mind:

1. A table should be formed in keeping with the objects of statistical enquiry.
2. A table should be carefully prepared so that it is easily understandable.
3. A table should be formed so as to suit the size of the paper. But such an adjustment should not be at the cost of legibility.
4. If the figures in the table are large, they should be suitably rounded or approximated. The method of approximation and units of measurements too should be specified.

5. Rows and columns in a table should be numbered and certain figures to be stressed may be put in ‘ box’ or ‘ circle’ or in bold letters.
6. The arrangements of rows and columns should be in a logical and systematic order. This arrangement may be alphabetical, chronological or according to size.
7. The rows and columns are separated by single, double or thick lines to represent various classes and sub-classes used. The corresponding proportions or percentages should be given in adjoining rows and columns to enable comparison. A vertical expansion of the table is generally more convenient than the horizontal one.
8. The averages or totals of different rows should be given at the right of the table and that of columns at the bottom of the table. Totals for every sub-class too should be mentioned.
9. In case it is not possible to accommodate all the information in a single table, it is better to have two or more related tables.

Type of Tables:

Tables can be classified according to their purpose, stage of enquiry, nature of data or number of characteristics used. On the basis of the number of characteristics, tables may be classified as follows:

1. Simple or one-way table 2. Two way table
3. Manifold table

Simple or one-way Table:

A simple or one-way table is the simplest table which contains data of one characteristic only. A simple table is easy to construct and simple to follow. For example, the blank table given below may be used to show the number of adults in different occupations in a locality.

The number of adults in different occupations in a locality

Occupations	No. Of Adults
Total	

Two-way Table:

A table, which contains data on two characteristics, is called a two-way table. In such case, therefore, either stub or caption is divided into two co-ordinate parts. In the given table, as an example the caption may be further divided in respect of ' sex' . This subdivision is shown in two-way table, which now contains two characteristics namely, occupation and sex.

The umber of adults in a locality in respect of occupation and sex

Occupation	**No. of Adults**		**Total**
	Male	**Female**	
Total			

Manifold Table:

Thus, more and more complex tables can be formed by including other characteristics. For example, we may further classify the caption sub-headings in the above table in respect of "marital status", " religion" and "socio-economic status" etc. A table ,which has more than two characteristics of data is considered as a manifold table. For instance , table shown below shows three characteristics namely, occupation, sex and marital status.

Occupation	No. of Adults						Total
	Male			Female			
	M	**U**	**Total**	**M**	**U**	**Total**	
Total							

Foot note: M Stands for Married and U stands for unmarried.

Manifold tables, though complex are good in practice as these enable full information to be incorporated and facilitate analysis of all related facts. Still, as a normal practice, not more than four characteristics should be represented in one table to avoid confusion. Other related tables may be formed to show the remaining characteristics

Exercise - 3

I. Choose the best answer:

1. When the collected data is grouped with reference to time, we have
   a) Quantitative classification b) Qualitative classification
   c) Geographical Classification d) Chorological Classification
2. Most quantitative classifications are
   a) Chronological b) Geographical
   c) Frequency Distribution d) None of these
3. Caption stands for
   a) A numerical information b) The column headings
   c) The row headings d) The table headings
4. A simple table contains data on
   a) Two characteristics b) Several characteristics
   c) One characteristic d) Three characteristics
5. The headings of the rows given in the first column of a table are called
   a) Stubs b) Captions
   c) Titles d) Reference notes

II. Fill in the blanks:

6. Geographical classification means, classification of data according to_____.
7. The data recorded according to standard of education like illiterate, primary, secondary, graduate, technical etc, will be known as_____classification.
8. An arrangement of data into rows and columns is known as _____.
9. Tabulation follows_____.
10. In a manifold table we have data on_____.

III. Answer the following questions:

11. Define three types of data.
12. Define primary and secondary data.
13. What are the points that are to be considered in the use of secondary data?
14. What are the sources of secondary data?
15. Give the merits and demerits of primary data.
16. State the characteristics of a good questionnaire.
17. Define classification.
18. What are the main objects of classification?
19. Write a detail note on the types of classification.
20. Define tabulation.
21. Give the advantages of tabulation.
22. What are the main parts of an ideal table? Explain.
23. What are the essential characteristics of a good table?
24. Define one-way and two-way table.
25. Explain manifold table with example.

IV. Suggested Activities:

26. Collect a primary data about the mode of transport of your school students. Classify the data and tabulate it.
27. Collect the important and relevant tables from various sources and include these in your album note book.

Answers:

1. (d) 2. (c) 3. (b) 4. (c) 5. (a)
6. Place
7. Qualitative
8. Tabulation
9. Classification
10. More than two characteristics

4. FREQUENCY DISTRIBUTION

Introduction:

Frequency distribution is a series when a number of observations with similar or closely related values are put in separate bunches or groups, each group being in order of magnitude in a series. It is simply a table in which the data are grouped into classes and the number of cases which fall in each class are recorded. It shows the frequency of occurrence of different values of a single Phenomenon.

A frequency distribution is constructed for three main reasons:

1. To facilitate the analysis of data.
2. To estimate frequencies of the unknown population distribution from the distribution of sample data and
3. To facilitate the computation of various statistical measures

Raw data:

The statistical data collected are generally raw data or ungrouped data. Let us consider the daily wages (in Rs) of 30 labourers in a factory.

80	70	55	50	60	65	40	30	80	90
75	45	35	65	70	80	82	55	65	80
60	55	38	65	75	85	90	65	45	75

The above figures are nothing but raw or ungrouped data and they are recorded as they occur without any pre consideration. This representation of data does not furnish any useful information and is rather confusing to mind. A better way to express the figures in an ascending or descending order of magnitude and is commonly known as array. But this does not reduce the bulk of the data. The above data when formed into an array is in the following form:

30	35	38	40	45	45	50	55	55	55
60	60	65	65	65	65	65	65	70	70
75	75	75	80	80	80	80	85	90	90

The array helps us to see at once the maximum and minimum values. It also gives a rough idea of the distribution of the items over the range . When we have a large number of items, the formation of an array is very difficult, tedious and cumbersome. The Condensation should be directed for better understanding and may be done in two ways, depending on the nature of the data.

a) Discrete (or) Ungrouped frequency distribution:

In this form of distribution, the frequency refers to discrete value. Here the data are presented in a way that exact measurement of units are clearly indicated.

There are definite difference between the variables of different groups of items. Each class is distinct and separate from the other class. Non-continuity from one class to another class exist. Data as such facts like the number of rooms in a house, the number of companies registered in a country, the number of children in a family, etc.

The process of preparing this type of distribution is very simple. We have just to count the number of times a particular value is repeated, which is called the frequency of that class. In order to facilitate counting prepare a column of tallies.

In another column, place all possible values of variable from the lowest to the highest. Then put a bar (Vertical line) opposite the particular value to which it relates.

To facilitate counting, blocks of five bars 𝍸 are prepared and some space is left in between each block. We finally count the number of bars and get frequency.

Example 1:

In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

1	0	3	2	1	5	6	2
2	1	0	3	4	2	1	6
3	2	1	5	3	3	2	4
2	2	3	0	2	1	4	5
3	3	4	4	1	2	4	5

Represent the data in the form of a discrete frequency distribution.

Solution:
Frequency distribution of the number of children

Number of Children	Tally Marks	Frequency
0	|||	3
1	𝍸 ||	7
2	𝍸 𝍸	10
3	𝍸 |||	8
4	𝍸 |	6
5	||||	4
6	||	2
	Total	40

b) Continuous frequency distribution:

In this form of distribution refers to groups of values. This becomes necessary in the case of some variables which can take any fractional value and in which case an exact measurement is not possible. Hence a discrete variable can be presented in the form of a continuous frequency distribution.

Wage distribution of 100 employees

Weekly wages (Rs)	Number of employees
50-100	4
100-150	12
150-200	22
200-250	33
250-300	16
300-350	8
350-400	5
Total	100

Nature of class:

The following are some basic technical terms when a continuous frequency distribution is formed or data are classified according to class intervals.

a) Class limits:

The class limits are the lowest and the highest values that can be included in the class. For example, take the class 30-40. The lowest value of the class is 30 and highest class is 40. The two boundaries of class are known as the lower limits and the upper limit of the class. The lower limit of a class is the value below which there can be no item in the class. The upper limit of a class is the value above which there can be no item to that class. Of the class 60-79, 60 is the lower limit and 79 is the upper limit, i.e. in the case there can be no value which is less than 60 or more than 79. The way in which class limits are stated depends upon the nature of the data. In statistical calculations, lower class limit is denoted by L and upper class limit by U.

b) Class Interval:

The class interval may be defined as the size of each grouping of data. For example, 50-75, 75-100, 100-125…are class intervals. Each grouping begins with the lower limit of a class interval and ends at the lower limit of the next succeeding class interval

c) Width or size of the class interval:

The difference between the lower and upper class limits is called Width or size of class interval and is denoted by ' C' .

d) Range:

The difference between largest and smallest value of the observation is called The Range and is denoted by ' R' ie

R = Largest value – Smallest value

R = L - S

e) Mid-value or mid-point:

The central point of a class interval is called the mid value or mid-point. It is found out by adding the upper and lower limits of a class and dividing the sum by 2.

(i.e.) Midvalue = $\frac{L + U}{2}$

For example, if the class interval is 20-30 then the mid-value is

$$\frac{20 + 30}{2} = 25$$

f) Frequency:

Number of observations falling within a particular class interval is called frequency of that class.

Let us consider the frequency distribution of weights if persons working in a company.

Weight (in kgs)	Number of persons
30-40	25
40-50	53
50-60	77
60-70	95
70-80	80
80-90	60
90-100	30
Total	420

In the above example, the class frequency are 25,53,77,95,80,60,30. The total frequency is equal to 420. The total frequency indicate the total number of observations considered in a frequency distribution.

g) Number of class intervals:

The number of class interval in a frequency is matter of importance. The number of class interval should not be too many. For an ideal frequency distribution, the number of class intervals can vary from 5 to 15. To decide the number of class intervals for the frequency distributive in the whole data, we choose the lowest and the highest of the values. The difference between them will enable us to decide the class intervals.

Thus the number of class intervals can be fixed arbitrarily keeping in view the nature of problem under study or it can be

decided with the help of Sturges' Rule. According to him, the number of classes can be determined by the formula

$$K = 1 + 3.322 \log_{10} N$$

Where N = Total number of observations

log = logarithm of the number

K = Number of class intervals.

Thus if the number of observation is 10, then the number of class intervals is

$$K = 1 + 3.322 \log 10 = 4.322 \cong 4$$

If 100 observations are being studied, the number of class interval is

$$K = 1 + 3.322 \log 100 = 7.644 \cong 8$$

and so on.

h) Size of the class interval:

Since the size of the class interval is inversely proportional to the number of class interval in a given distribution. The approximate value of the size (or width or magnitude) of the class interval ' C' is obtained by using sturges rule as

$$\text{Size of class interval} = C = \frac{\text{Range}}{\text{Number of class interval}}$$

$$= \frac{\text{Range}}{1+3.322 \log_{10} N}$$

Where Range = Largest Value – smallest value in the distribution.

Types of class intervals:

There are three methods of classifying the data according to class intervals namely

a) Exclusive method
b) Inclusive method
c) Open-end classes

a) Exclusive method:

When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class; it is known as the exclusive method of classification. The following data are classified on this basis.

Expenditure (Rs.)	No. of families
0 - 5000	60
5000-10000	95
10000-15000	122
15000-20000	83
20000-25000	40
Total	400

It is clear that the exclusive method ensures continuity of data as much as the upper limit of one class is the lower limit of the next class. In the above example, there are so families whose expenditure is between Rs.0 and Rs.4999.99. A family whose expenditure is Rs.5000 would be included in the class interval 5000-10000. This method is widely used in practice.

b) Inclusive method:

In this method, the overlapping of the class intervals is avoided. Both the lower and upper limits are included in the class interval. This type of classification may be used for a grouped frequency distribution for discrete variable like members in a family, number of workers in a factory etc., where the variable may take only integral values. It cannot be used with fractional values like age, height, weight etc.

This method may be illustrated as follows:

Class interval	Frequency
5- 9	7
10-14	12
15-19	15
20-29	21
30-34	10
35-39	5
Total	70

Thus to decide whether to use the inclusive method or the exclusive method, it is important to determine whether the variable

under observation in a continuous or discrete one. In case of continuous variables, the exclusive method must be used. The inclusive method should be used in case of discrete variable.

c) Open end classes:

A class limit is missing either at the lower end of the first class interval or at the upper end of the last class interval or both are not specified. The necessity of open end classes arises in a number of practical situations, particularly relating to economic and medical data when there are few very high values or few very low values which are far apart from the majority of observations.

The example for the open-end classes as follows :

Salary Range	**No of workers**
Below 2000	7
2000 – 4000	5
4000 – 6000	6
6000 – 8000	4
8000 and above	3

Construction of frequency table:

Constructing a frequency distribution depends on the nature of the given data. Hence, the following general consideration may be borne in mind for ensuring meaningful classification of data.

1. The number of classes should preferably be between 5 and 20. However there is no rigidity about it.
2. As far as possible one should avoid values of class intervals as 3,7,11,26..etc. preferably one should have class-intervals of either five or multiples of 5 like 10,20,25,100 etc.
3. The starting point i.e the lower limit of the first class, should either be zero or 5 or multiple of 5.
4. To ensure continuity and to get correct class interval we should adopt “exclusive” method.
5. Wherever possible, it is desirable to use class interval of equal sizes.

Preparation of frequency table:

The premise of data in the form of frequency distribution describes the basic pattern which the data assumes in the mass. Frequency distribution gives a better picture of the pattern of data if the number of items is large. If the identity of the individuals about whom a particular information is taken, is not relevant then the first step of condensation is to divide the observed range of variable into a suitable number of class-intervals and to record the number of observations in each class. Let us consider the weights in kg of 50 college students.

42	62	46	54	41	37	54	44	32	45
47	50	58	49	51	42	46	37	42	39
54	39	51	58	47	64	43	48	49	48
49	61	41	40	58	49	59	57	57	34
56	38	45	52	46	40	63	41	51	41

Here the size of the class interval as per sturges rule is obtained as follows

$$\text{Size of class interval} = C = \frac{\text{Range}}{1+3.322 \log N}$$

$$= \frac{64 - 32}{1+3.322 \log(50)} = \frac{32}{6.64} \quad 5$$

Thus the number of class interval is 7 and size of each class is 5. The required size of each class is 5. The required frequency distribution is prepared using tally marks as given below:

Class Interval	Tally marks	Frequency
30-35	||	2
35-40	𝍸 |	6
40-45	𝍸 𝍸 ||	12
45-50	𝍸 𝍸 ||||	14
50-55	𝍸 |	6
55-60	𝍸 |	6
60-65	||||	4
Total		50

Example 2:
Given below are the number of tools produced by workers in a factory.

43	18	25	18	39	44	19	20	20	26
40	45	38	25	13	14	27	41	42	17
34	31	32	27	33	37	25	26	32	25
33	34	35	46	29	34	31	34	35	24
28	30	41	32	29	28	30	31	30	34
31	35	36	29	26	32	36	35	36	37
32	23	22	29	33	37	33	27	24	36
23	42	29	37	29	23	44	41	45	39
21	21	42	22	28	22	15	16	17	28
22	29	35	31	27	40	23	32	40	37

Construct frequency distribution with inclusive type of class interval. Also find.

1. How many workers produced more than 38 tools?
2. How many workers produced less than 23 tools?

Solution:
Using sturges formula for determining the number of class intervals, we have

Number of class intervals = $1+ 3.322 \log_{10}N$

$= 1+ 3.322 \log_{10}100$

$= 7.6$

$$\text{Sizes of class interval} = \frac{\text{Range}}{\text{Number of class interval}}$$

$$= \frac{46 - 13}{7.6}$$

$$5$$

Hence taking the magnitude of class intervals as 5, we have 7 classes 13-17, 18-22… 43-47 are the classes by inclusive type. Using tally marks, the required frequency distribution is obtain in the following table

Class Interval	Tally Marks	Number of tools produced (Frequency)
13-17	𝍸 |	6
18-22	𝍸 𝍸 |	11
23-27	𝍸 𝍸 𝍸 |||	18
28-32	𝍸 𝍸 𝍸 𝍸 𝍸	25
33-37	𝍸 𝍸 𝍸 𝍸 ||	22
38-42	𝍸 𝍸 |	11
43-47	𝍸 ||	7
Total		100

Percentage frequency table:

The comparison becomes difficult and at times impossible when the total number of items are large and highly different one distribution to other. Under these circumstances percentage frequency distribution facilitates easy comparability. In percentage frequency table, we have to convert the actual frequencies into percentages. The percentages are calculated by using the formula given below:

$$\text{Frequency percentage} = \frac{\text{Actual Frequency}}{\text{Total Frequency}} \times 100$$

It is also called relative frequency table:

An example is given below to construct a percentage frequency table.

Marks	No. of students	Frequency percentage
0-10	3	6
10-20	8	16
20-30	12	24
30-40	17	34
40-50	6	12
50-60	4	8
Total	50	100

Cumulative frequency table:

Cumulative frequency distribution has a running total of the values. It is constructed by adding the frequency of the first class interval to the frequency of the second class interval. Again add that total to the frequency in the third class interval continuing until the final total appearing opposite to the last class interval will be the total of all frequencies. The cumulative frequency may be downward or upward. A downward cumulation results in a list presenting the number of frequencies "less than" any given amount as revealed by the lower limit of succeeding class interval and the upward cumulative results in a list presenting the number of frequencies "more than" and given amount is revealed by the upper limit of a preceding class interval.

Example 3:

Age group (in years)	Number of women	Less than Cumulative frequency	More than cumulative frequency
15-20	3	3	64
20-25	7	10	61
25-30	15	25	54
30-35	21	46	39
35-40	12	58	18
40-45	6	64	6

(a) Less than cumulative frequency distribution table

End values upper limit	less than Cumulative frequency
Less than 20	3
Less than 25	10
Less than 30	25
Less than 35	46
Less than 40	58
Less than 45	64

(b) More than cumulative frequency distribution table

End values lower limit	Cumulative frequency more than
15 and above	64
20 and above	61
25 and above	54
30 and above	39
35 and above	18
40 and above	6

4.8.1 Conversion of cumulative frequency to simple Frequency:

If we have only cumulative frequency ‘ either less than or more than’ , we can convert it into simple frequencies. For example if we have ‘ less than Cumulative frequency, we can convert this to simple frequency by the method given below:

Class interval	‘ less than’ Cumulative frequency	Simple frequency
15-20	3	3
20-25	10	10 – 3 = 7
25-30	25	25 – 10 = 15
30-35	46	46 – 25 = 21
35-40	58	58 – 46 = 12
40-45	64	64 – 58 = 6

Method of converting ‘ more than’ cumulative frequency to simple frequency is given below.

Class interval	‘ more than’ Cumulative frequency	Simple frequency
15-20	64	64 – 61 = 3
20-25	61	61 – 54 = 7
25-30	54	54 –39 = 15
30-35	39	39 – 18 = 21
35-40	18	18 – 6 = 12
40-45	6	6 – 0 = 6

Cumulative percentage Frequency table:

Instead of cumulative frequency, if cumulative percentages are given, the distribution is called cumulative percentage frequency distribution. We can form this table either by converting the frequencies into percentages and then cumulate it or we can convert the given cumulative frequency into percentages.

Example 4:

Income (in Rs)	No. of family	Cumulative frequency	Cumulative percentage
2000-4000	8	8	5.7
4000-6000	15	23	16.4
6000-8000	27	50	35.7
8000-10000	44	94	67.1
10000-12000	31	125	89.3
12000-14000	12	137	97.9
14000-20000	3	140	100.0
Total	140		

Bivariate frequency distribution:

In the previous sections, we described frequency distribution involving one variable only. Such frequency distributions are called univariate frequency distribution. In many situations simultaneous study of two variables become necessary. For example, we want to classify data relating to the weights are height of a group of individuals, income and expenditure of a group of individuals, age of husbands and wives.

The data so classified on the basis of two variables give rise to the so called bivariate frequency distribution and it can be summarized in the form of a table is called bivariate (two-way) frequency table. While preparing a bivariate frequency distribution, the values of each variable are grouped into various classes (not necessarily the same for each variable) . If the data corresponding to one variable, say X is grouped into m classes and the data corresponding to the other variable, say Y is grouped into n classes then the bivariate table will consist of mxn cells. By going through the different pairs of the values, (X,Y) of the variables and using tally marks we can find the frequency of each

cell and thus, obtain the bivariate frequency table. The formate of a bivariate frequency table is given below:

Format of Bivariate Frequency table

x-series / y-series		Class-Intervals Mid-values	Marginal Frequency of Y
Class-intervals	MidValues		f_y
Marginal frequency of X		f_x	Total $\Sigma f_x = \Sigma f_y = N$

Here f(x,y) is the frequency of the pair (x,y). The frequency distribution of the values of the variables x together with their frequency total (fx) is called the marginal distribution of x and the frequency distribution of the values of the variable Y together with the total frequencies is known as the marginal frequency distribution of Y. The total of the values of manual frequencies is called grand total (N)

Example 5:

The data given below relate to the height and weight of 20 persons. Construct a bivariate frequency table with class interval of height as 62-64, 64-66…and weight as 115-125,125-135, write down the marginal distribution of X and Y.

S.No.	Height	Weight	S.No.	Height	Weight
1	70	170	11	70	163
2	65	135	12	67	139
3	65	136	13	63	122
4	64	137	14	68	134
5	69	148	15	67	140
6	63	121	16	69	132
7	65	117	17	65	120
8	70	128	18	68	148
9	71	143	19	67	129
10	62	129	20	67	152

Solution:

Bivariate frequency table showing height and weight of persons.

Height(x) / Weight(y)	62-64	64-66	66-68	68-70	70-72	Total
115-125	II (2)	II (2)				4
125-135	I (1)		I (1)	II (2)	I (1)	5
135-145		III (3)	II (2)		I (1)	6
145-155			I (1)	II (2)		3
155-165					I (1)	1
165-175					I (1)	1
Total	3	5	4	4	4	20

The marginal distribution of height and weight are given in the following table.

Marginal distribution of height (X)		Marginal distribution of (Y)	
CI	Frequency	CI	Frequency
62-64	3	115-125	4
64-66	5	125-135	5
66-68	4	135-145	6
68-70	4	145-155	3
70-72	4	155-165	1
Total	20	165-175	1
		Total	20

Exercise - 4

I. Choose the best answer:

1. In an exclusive class interval
   (a) the upper class limit is exclusive.
   (b) the lower class limit is exclusive.
   (c) the lower and upper class limits are exclusive.
   (d) none of the above.
2. If the lower and upper limits of a class are 10 and 40 respectively, the mid points of the class is
   (a) 15.0 (b) 12.5 (c) 25.0 (d) 30.0
3. Class intervals of the type 30-39,40-49,50-59 represents
   (a) inclusive type (b) exclusive type
   (c) open-end type (d) none.
4. The class interval of the continuous grouped data is

10-19	20-29	30-39	40-49	50-59

   (a) 9 (b)10 (c) 14.5 (d) 4.5
5. Raw data means
   (a) primary data (b) secondary data
   (c) data collected for investigation (d)Well classified data.

II. Fill in the blanks:

6. H.A.Sturges formula for finding number of classes is _____.
7. If the mid-value of a class interval is 20 and the difference between two consecutive midvalues is 10 the class limits are_____and_____.
8. The difference between the upper and lower limit of class is called_____.
9. The average of the upper and lower limits of a class is known as_____.
10. Number of observations falling within a particular class interval is called_____of that class.

III. Answer the following questions:

11. What is a frequency distribution?
12. What is an array?
13. What is discrete and continuous frequency distribution?

14. Distinguish between with suitable example.
    (i) Continuous and discrete frequency
    (ii) Exclusive and Inclusive class interval
    (iii) Less than and more than frequency table
    (iv) Simple and Bivariate frequency table.
15. The following data gives the number of children in 50 families. Construct a discrete frequency table.

4	2	0	2	3	2	2	1	0	2
3	5	1	1	4	2	1	3	4	2
6	1	2	2	2	1	3	4	1	0
1	3	4	1	0	1	2	2	2	5
2	4	3	0	1	3	6	1	0	1

16. In a survey, it was found that 64 families bought milk in the following quantities in a particular month. Quantity of milk (in litres) bought by 64 Families in a month. Construct a continuous frequency distribution making classes of 5-9, 10-14 and so on.

19	16	22	9	22	12	39	19
14	23	6	24	16	18	7	17
20	25	28	18	10	24	20	21
10	7	18	28	24	20	14	23
25	34	22	5	33	23	26	29
13	36	11	26	11	37	30	13
8	15	22	21	32	21	31	17
16	23	12	9	15	27	17	21

17. 25 values of two variables X and Y are given below. Form a two-way frequency table showing the relationship between the two. Take class interval of X as 10-20,20-30,…. and Y as 100-200,200-300,….

X	Y	X	Y	X	Y
12	140	36	315	57	416
24	256	27	440	44	380
33	360	57	390	48	492
22	470	21	590	48	370
44	470	51	250	52	312
37	380	27	550	41	330
29	280	42	360	69	590
55	420	43	570		
48	390	52	290		

18. The ages of 20 husbands and wives are given below. Form a two-way frequency table on the basis of ages of husbands and wives with the class intervals 20-25,25-30 etc.

Age of husband	Age of wife
28	23
37	30
42	40
25	26
29	25
47	41
37	35
35	25
23	21
41	38

Age of husband	Age of wife
27	24
39	34
23	20
33	31
36	29
32	35
22	23
29	29
38	34
48	47

IV .Suggested Activities:

From the mark sheets of your class, form the frequency tables, less than and more than cumulative frequency tables.

Answers

I. 1. (a) 2. (c) 3.(a) 4.(b) 5. (a)

II. 6. $k = 1 + 3.322 \log_{10} N$ 7. 15 and 25

8. width or size of class 9. Mid-value 10. Frequency

5. DIAGRAMATIC AND GRAPHICAL REPRESENTATION

Introduction:

In the previous chapter, we have discussed the techniques of classification and tabulation that help in summarising the collected data and presenting them in a systematic manner. However, these forms of presentation do not always prove to be interesting to the common man. One of the most convincing and appealing ways in which statistical results may be presented is through diagrams and graphs. Just one diagram is enough to represent a given data more effectively than thousand words.

Moreover even a layman who has nothing to do with numbers can also understands diagrams. Evidence of this can be found in newspapers, magazines, journals, advertisement, etc. An attempt is made in this chapter to illustrate some of the major types of diagrams and graphs frequently used in presenting statistical data.

Diagrams:

A diagram is a visual form for presentation of statistical data, highlighting their basic facts and relationship. If we draw diagrams on the basis of the data collected they will easily be understood and appreciated by all. It is readily intelligible and save a considerable amount of time and energy.

Significance of Diagrams and Graphs:

Diagrams and graphs are extremely useful because of the following reasons.

1. They are attractive and impressive.
2. They make data simple and intelligible.
3. They make comparison possible
4. They save time and labour.
5. They have universal utility.
6. They give more information.
7. They have a great memorizing effect.

General rules for constructing diagrams:

The construction of diagrams is an art, which can be acquired through practice. However, observance of some general guidelines can help in making them more attractive and effective. The diagrammatic presentation of statistical facts will be advantageous provided the following rules are observed in drawing diagrams.

1. A diagram should be neatly drawn and attractive.
2. The measurements of geometrical figures used in diagram should be accurate and proportional.
3. The size of the diagrams should match the size of the paper.
4. Every diagram must have a suitable but short heading.
5. The scale should be mentioned in the diagram.
6. Diagrams should be neatly as well as accurately drawn with the help of drawing instruments.
7. Index must be given for identification so that the reader can easily make out the meaning of the diagram.
8. Footnote must be given at the bottom of the diagram.
9. Economy in cost and energy should be exercised in drawing diagram.

Types of diagrams:

In practice, a very large variety of diagrams are in use and new ones are constantly being added. For the sake of convenience and simplicity, they may be divided under the following heads:

1. One-dimensional diagrams
2. Two-dimensional diagrams
3. Three-dimensional diagrams
4. Pictograms and Cartograms

One-dimensional diagrams:

In such diagrams, only one-dimensional measurement, i.e height is used and the width is not considered. These diagrams are in the form of bar or line charts and can be classified as

1. Line Diagram
2. Simple Diagram
3. Multiple Bar Diagram
4. Sub-divided Bar Diagram
5. Percentage Bar Diagram

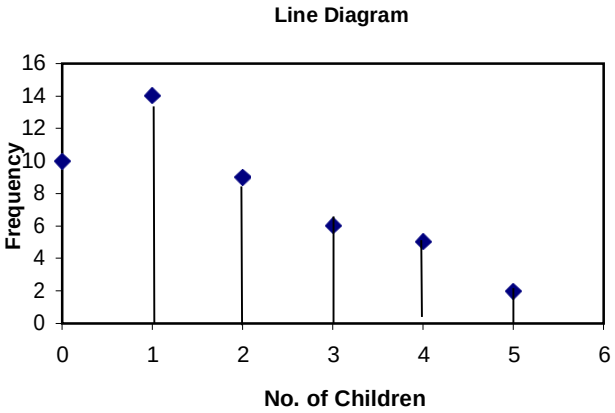
Line Diagram:

Line diagram is used in case where there are many items to be shown and there is not much of difference in their values. Such diagram is prepared by drawing a vertical line for each item according to the scale. The distance between lines is kept uniform. Line diagram makes comparison easy, but it is less attractive.

Example 1:

Show the following data by a line chart:

No. of children	0	1	2	3	4	5
Frequency	10	14	9	6	4	2



Simple Bar Diagram:

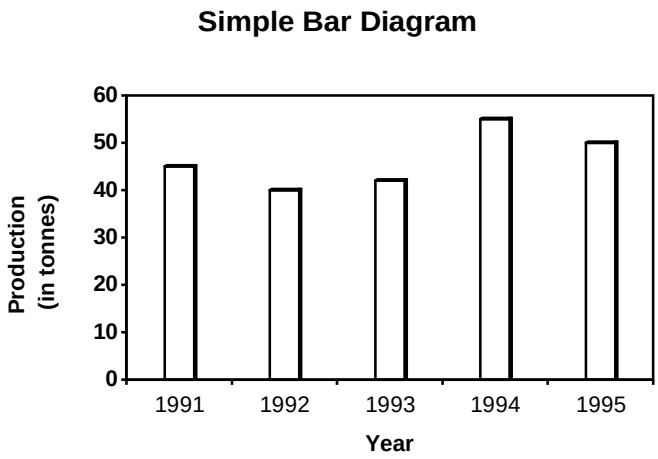
Simple bar diagram can be drawn either on horizontal or vertical base, but bars on horizontal base more common. Bars must be uniform width and intervening space between bars must be equal.While constructing a simple bar diagram, the scale is determined on the basis of the highest value in the series.

To make the diagram attractive, the bars can be coloured. Bar diagram are used in business and economics. However, an important limitation of such diagrams is that they can present only one classification or one category of data. For example, while presenting the population for the last five decades, one can only depict the total population in the simple bar diagrams, and not its sex-wise distribution.

Example 2:
Represent the following data by a bar diagram.

Year	Production (in tones)
1991	45
1992	40
1993	42
1994	55
1995	50

Solution:



Multiple Bar Diagram:

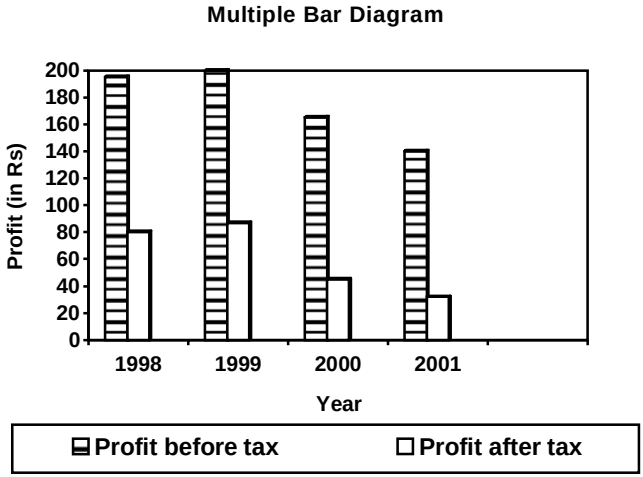
Multiple bar diagram is used for comparing two or more sets of statistical data. Bars are constructed side by side to represent the set of values for comparison. In order to distinguish bars, they may be either differently coloured or there should be different types of crossings or dotting, etc. An index is also prepared to identify the meaning of different colours or dottings.

Example 3:

Draw a multiple bar diagram for the following data.

Year	Profit before tax (in lakhs of rupees)	Profit after tax (in lakhs of rupees)
1998	195	80
1999	200	87
2000	165	45
2001	140	32

Solution:



Sub-divided Bar Diagram:

In a sub-divided bar diagram, the bar is sub-divided into various parts in proportion to the values given in the data and the whole bar represent the total. Such diagrams are also called Component Bar diagrams. The sub divisions are distinguished by different colours or crossings or dottings.

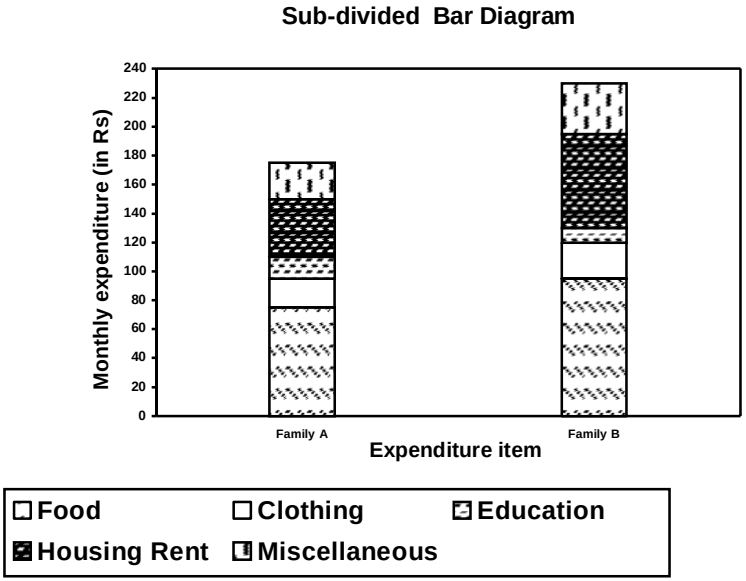
The main defect of such a diagram is that all the parts do not have a common base to enable one to compare accurately the various components of the data.

Example 4:

Represent the following data by a sub-divided bar diagram.

Expenditure items	Monthly expenditure (in Rs.)	
	Family A	Family B
Food	75	95
Clothing	20	25
Education	15	10
Housing Rent	40	65
Miscellaneous	25	35

Solution:



Percentage bar diagram:

This is another form of component bar diagram. Here the components are not the actual values but percentages of the whole. The main difference between the sub-divided bar diagram and percentage bar diagram is that in the former the bars are of different heights since their totals may be different whereas in the latter the bars are of equal height since each bar represents 100 percent. In the case of data having sub-division, percentage bar diagram will be more appealing than sub-divided bar diagram.

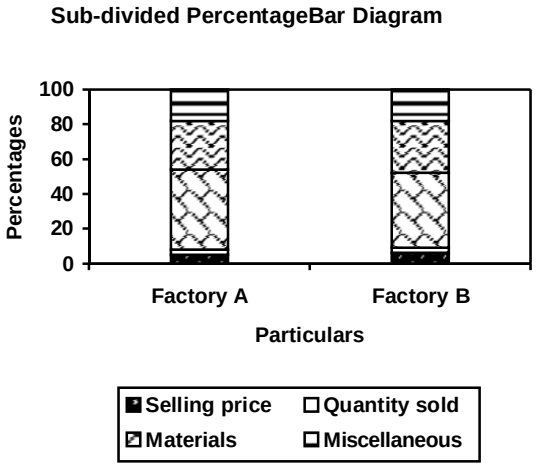
Example 5:
Represent the following data by a percentage bar diagram.

Particular	Factory A	Factory B
Selling Price	400	650
Quantity Sold	240	365
Wages	3500	5000
Materials	2100	3500
Miscellaneous	1400	2100

Solution:
Convert the given values into percentages as follows:

Particulars	Factory A		Factory B	
	Rs.	%	Rs.	%
Selling Price	400	5	650	6
Quantity Sold	240	3	365	3
Wages	3500	46	5000	43
Materials	2100	28	3500	30
Miscellaneous	1400	18	2100	18
Total	7640	100	11615	100

Solution:



Two-dimensional Diagrams:

In one-dimensional diagrams, only length 9 is taken into account. But in two-dimensional diagrams the area represent the data and so the length and breadth have both to be taken into account. Such diagrams are also called area diagrams or surface diagrams. The important types of area diagrams are:

1. Rectangles 2. Squares 3. Pie-diagrams

Rectangles:

Rectangles are used to represent the relative magnitude of two or more values. The area of the rectangles are kept in proportion to the values. Rectangles are placed side by side for comparison. When two sets of figures are to be represented by rectangles, either of the two methods may be adopted.

We may represent the figures as they are given or may convert them to percentages and then subdivide the length into various components. Thus the percentage sub-divided rectangular diagram is more popular than sub-divided rectangular since it enables comparison to be made on a percentage basis.

Example 6:

Represent the following data by sub-divided percentage rectangular diagram.

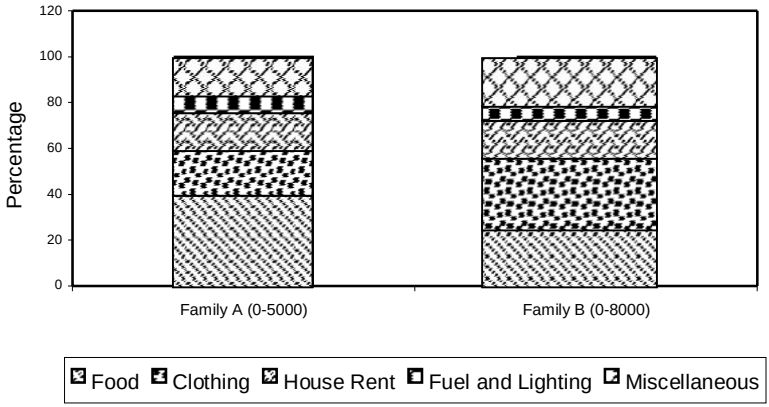
Items of Expenditure	Family A (Income Rs.5000)	Family B (income Rs.8000)
Food	2000	2500
Clothing	1000	2000
House Rent	800	1000
Fuel and lighting	400	500
Miscellaneous	800	2000
Total	5000	8000

Solution:

The items of expenditure will be converted into percentage as shown below:

Items of Expenditure	Family A		Family B	
	Rs.	Y	Rs.	Y
Food	2000	40	2500	31
Clothing	1000	20	2000	25
House Rent	800	16	1000	13
Fuel and Lighting	400	8	500	6
Miscellaneous	800	16	2000	25
Total	5000	100	8000	100

SUBDIVIDED PERCENTAGE RECTANGULAR DIAGRAM



Squares:

The rectangular method of diagrammatic presentation is difficult to use where the values of items vary widely. The method of drawing a square diagram is very simple. One has to take the square root of the values of various item that are to be shown in the diagrams and then select a suitable scale to draw the squares.

Example 7:

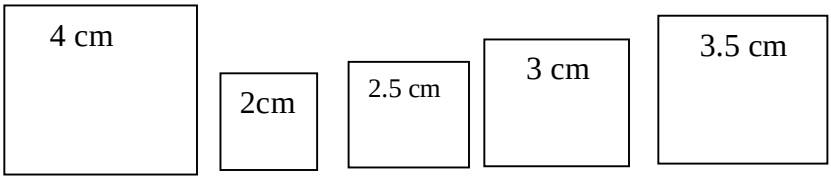
Yield of rice in Kgs. per acre of five countries are

Country	U.S.A	Australia	U.K	Canada	India
Yield of rice in Kgs per acre	6400	1600	2500	3600	4900

Represent the above data by Square diagram.

Solution: To draw the square diagram we calculate as follows:

Country	Yield	Square root	Side of the square in cm
U.S.A	6400	80	4
Australia	1600	40	2
U.K.	2500	50	2.5
Canada	3600	60	3
India	4900	70	3.5



Pie Diagram or Circular Diagram:

Another way of preparing a two-dimensional diagram is in the form of circles. In such diagrams, both the total and the component parts or sectors can be shown. The area of a circle is proportional to the square of its radius.

While making comparisons, pie diagrams should be used on a percentage basis and not on an absolute basis. In constructing a pie diagram the first step is to prepare the data so that various components values can be transposed into corresponding degrees on the circle.

The second step is to draw a circle of appropriate size with a compass. The size of the radius depends upon the available space and other factors of presentation. The third step is to measure points on the circle and representing the size of each sector with the help of a protractor.

Example 8:

Draw a Pie diagram for the following data of production of sugar in quintals of various countries.

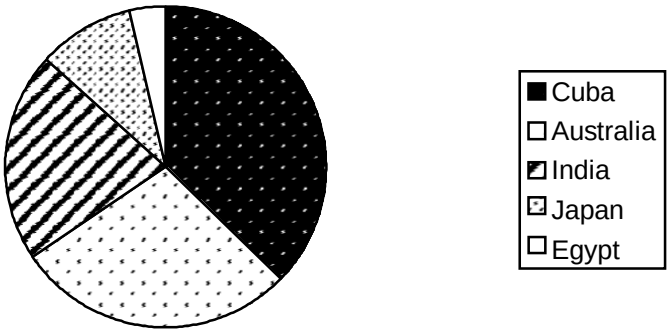
Country	Production of Sugar (in quintals)
Cuba	62
Australia	47
India	35
Japan	16
Egypt	6

Solution:

The values are expressed in terms of degree as follows.

Country	Production of Sugar	
	In Quintals	In Degrees
Cuba	62	134
Australia	47	102
India	35	76
Japan	16	35
Egypt	6	13
Total	166	360

Pie Diagram



Three-dimensional diagrams:

Three-dimensional diagrams, also known as volume diagram, consist of cubes, cylinders, spheres, etc. In such diagrams three things, namely length, width and height have to be taken into account. Of all the figures, making of cubes is easy. Side of a cube is drawn in proportion to the cube root of the magnitude of data.

Cubes of figures can be ascertained with the help of logarithms. The logarithm of the figures can be divided by 3 and the antilog of that value will be the cube-root.

Example 9:

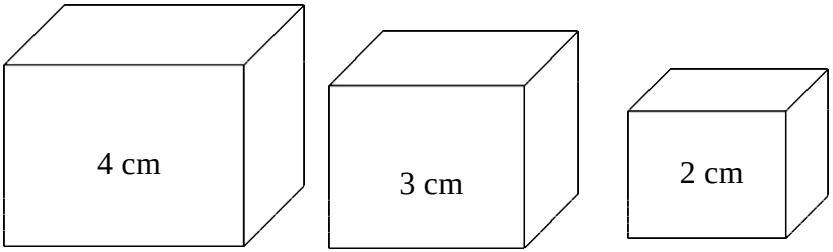
Represent the following data by volume diagram.

Category	Number of Students
Under graduate	64000
Post graduate	27000
Professionals	8000

Solution:

The sides of cubes can be determined as follows

Category	Number of students	Cube root	Side of cube
Undergraduate	64000	40	4 cm
Postgraduate	27000	30	3 cm
Professional	8000	20	2 cm



Undergraduate **Postgraduate** **professional**

Pictograms and Cartograms:

Pictograms are not abstract presentation such as lines or bars but really depict the kind of data we are dealing with. Pictures are attractive and easy to comprehend and as such this method is particularly useful in presenting statistics to the layman. When Pictograms are used, data are represented through a pictorial symbol that is carefully selected.

Cartograms or statistical maps are used to give quantitative information as a geographical basis. They are used to represent

spatial distributions. The quantities on the map can be shown in many ways such as through shades or colours or dots or placing pictogram in each geographical unit.

Graphs:

A graph is a visual form of presentation of statistical data. A graph is more attractive than a table of figure. Even a common man can understand the message of data from the graph. Comparisons can be made between two or more phenomena very easily with the help of a graph.

However here we shall discuss only some important types of graphs which are more popular and they are

1. Histogram 2. Frequency Polygon
3. Frequency Curve 4. Ogive 5. Lorenz Curve

5.6.1 Histogram:

A histogram is a bar chart or graph showing the frequency of occurrence of each value of the variable being analysed. In histogram, data are plotted as a series of rectangles. Class intervals are shown on the ' X-axis' and the frequencies on the ' Y-axis'.

The height of each rectangle represents the frequency of the class interval. Each rectangle is formed with the other so as to give a continuous picture. Such a graph is also called staircase or block diagram.

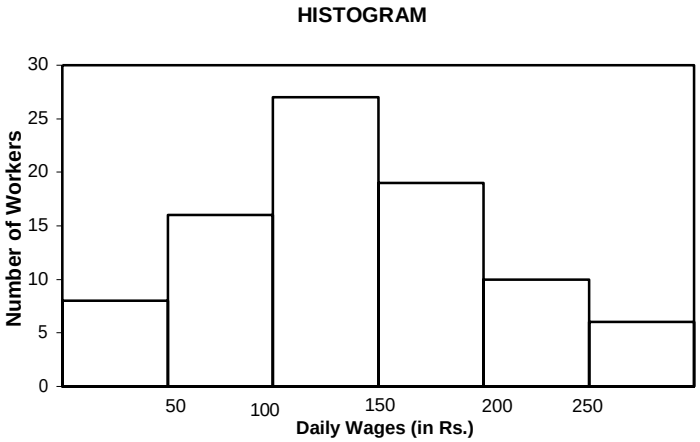
However, we cannot construct a histogram for distribution with open-end classes. It is also quite misleading if the distribution has unequal intervals and suitable adjustments in frequencies are not made.

Example 10:

Draw a histogram for the following data.

Daily Wages	Number of Workers
0-50	8
50-100	16
100-150	27
150-200	19
200-250	10
250-300	6

Solution:



Example 11:

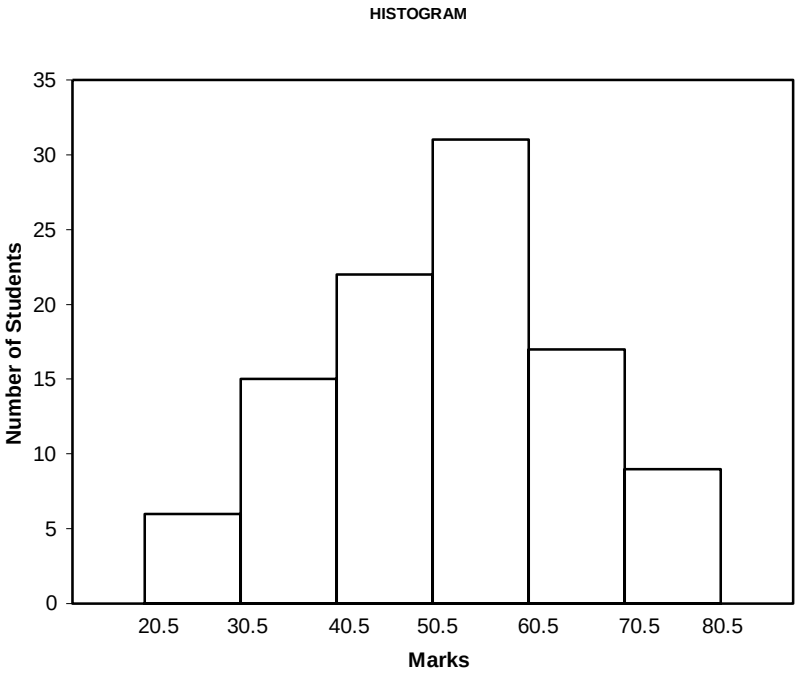
For the following data, draw a histogram.

Marks	Number of Students
21-30	6
31-40	15
41-50	22
51-60	31
61-70	17
71-80	9

Solution:

For drawing a histogram, the frequency distribution should be continuous. If it is not continuous, then first make it continuous as follows.

Marks	Number of Students
20.5-30.5	6
30.5-40.5	15
40.5-50.5	22
50.5-60.5	31
60.5-70.5	17
70.5-80.5	9



Example 12:

Draw a histogram for the following data.

Profits (in lakhs)	Number of Companies
0-10	4
10-20	12
20-30	24
30-50	32
50-80	18
80-90	9
90-100	3

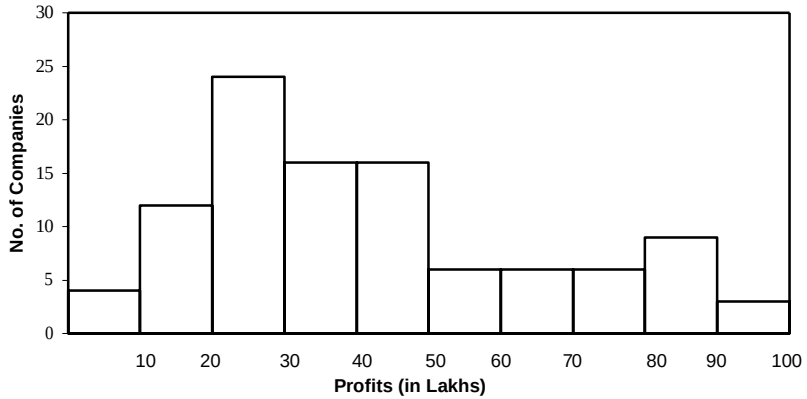
Solution:

When the class intervals are unequal, a correction for unequal class intervals must be made. The frequencies are adjusted as follows: The frequency of the class 30-50 shall be divided by two since the class interval is in double. Similarly the class interval 50-80 can be divided by 3. Then draw the histogram.

Now we rewrite the frequency table as follows.

Profits (in lakhs)	Number of Companies
0-10	4
10-20	12
20-30	24
30-40	16
40-50	16
50-60	6
60-70	6
70-80	6
80-90	9
90-100	3

HISTOGRAM

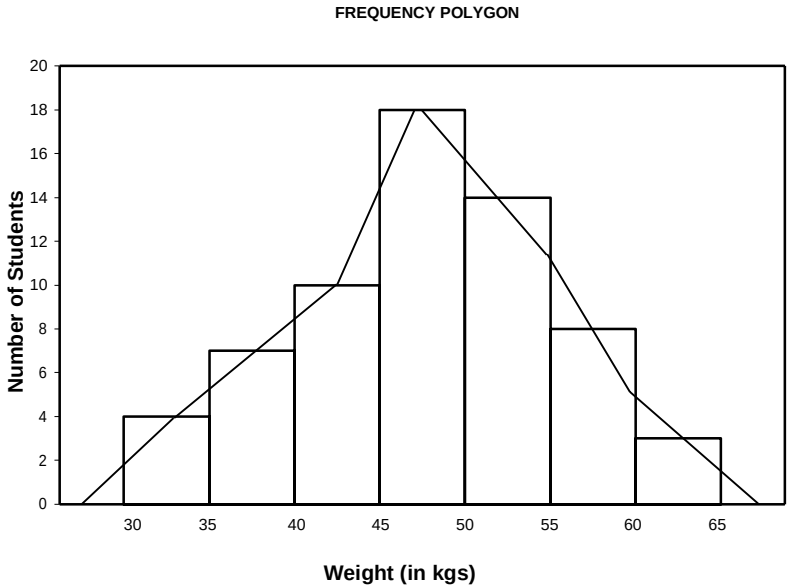


5.6.2 Frequency Polygon:

If we mark the midpoints of the top horizontal sides of the rectangles in a histogram and join them by a straight line, the figure so formed is called a Frequency Polygon. This is done under the assumption that the frequencies in a class interval are evenly distributed throughout the class. The area of the polygon is equal to the area of the histogram, because the area left outside is just equal to the area included in it.

Example 13:
Draw a frequency polygon for the following data.

Weight (in kg)	Number of Students
30-35	4
35-40	7
40-45	10
45-50	18
50-55	14
55-60	8
60-65	3



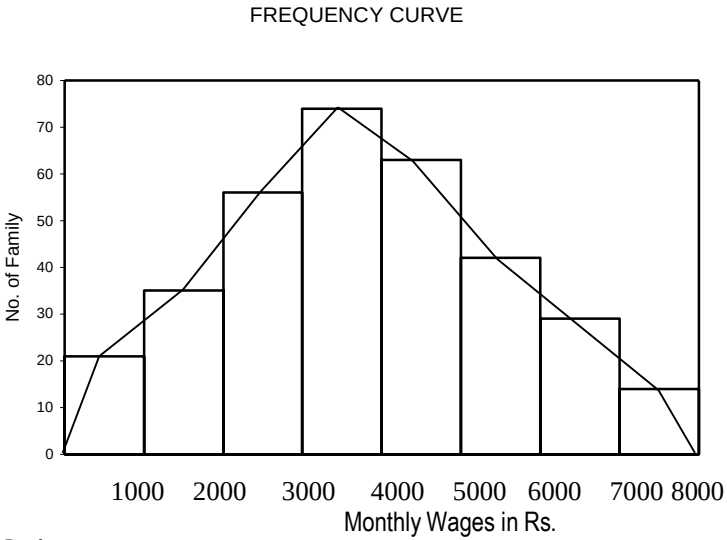
Frequency Curve:

If the middle point of the upper boundaries of the rectangles of a histogram is corrected by a smooth freehand curve, then that diagram is called frequency curve. The curve should begin and end at the base line.

Example 14:
Draw a frequency curve for the following data.

Monthly Wages (in Rs.)	No. of family
0-1000	21
1000-2000	35
2000-3000	56
3000-4000	74
4000-5000	63
5000-6000	40
6000-7000	29
7000-8000	14

Solution:



Ogives:

For a set of observations, we know how to construct a frequency distribution. In some cases we may require the number of observations less than a given value or more than a given value. This is obtained by a accumulating (adding) the frequencies upto

(or above) the give value. This accumulated frequency is called cumulative frequency.

These cumulative frequencies are then listed in a table is called cumulative frequency table. The curve table is obtained by plotting cumulative frequencies is called a cumulative frequency curve or an ogive.

There are two methods of constructing ogive namely:

1. The ‘ less than ogive’ method
2. The ‘ more than ogive’ method.

In less than ogive method we start with the upper limits of the classes and go adding the frequencies. When these frequencies are plotted, we get a rising curve. In more than ogive method, we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class. When these frequencies are plotted we get a declining curve.

Example 15:

Draw the Ogives for the following data.

Class interval	Frequency
20-30	4
30-40	6
40-50	13
50-60	25
60-70	32
70-80	19
80-90	8
90-100	3

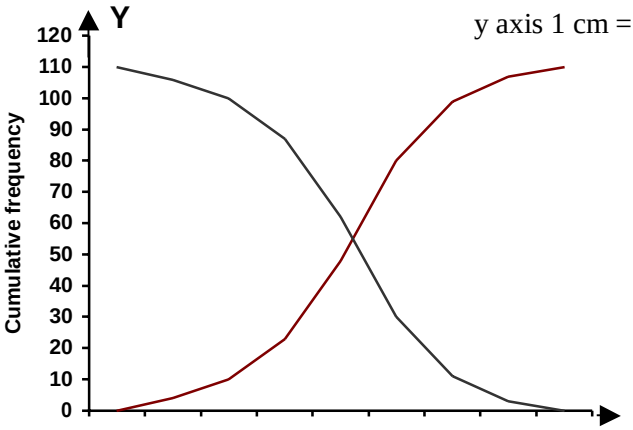
Solution:

Class limit	Less than ogive	More than ogive
20	0	110
30	4	106
40	10	100
50	23	87
60	48	62
70	80	30
80	99	11

90	107	3
100	110	0

Ogives

x axis 1cm = 10 units
y axis 1 cm = 10 units



20 30 40 50 60 70 80 90 100 X

Class limit

I Choose the best answer:

1. Which of the following is one dimensional diagram.
   (a) Bar diagram (b) Pie diagram (c) Cylinder
   (d) Histogram
2. Percentage bar diagram has
   (a) data expressed in percentages
   (b) equal width
   (c) equal interval
   (d) equal width and equal interval
3. Frequency curve
   (a) begins at the origin (b) passes through the origin
   (c) begins at the horizontal line.
   (d) begins and ends at the base line.
4. With the help of histogram we can draw
   (a) frequency polygon (b) frequency curve
   (c) frequency distribution
   (d) all the above
5. Ogives for more than type and less than type distribution intersect at
   (a) mean (b) median
   (c) mode (d) origin

II Fill in the blanks:

1. Sub-divided bar diagram are also called_____diagram.
2. In rectangular diagram, comparison is based on_____of the rectangles.
3. Squares are_____dimensional diagrams.
4. Ogives for more than type and less than type distribution intersects at_____.
5. _____Curve is graphical method of studying dispersion.

1. Represent the following data by a bar diagram.

Year	Profit (in thousands)
1995	2
1996	6
1997	11
1998	15
1999	20
2000	27

2. Represent the following data by a multiple bar diagram.

Factory	Workers	
	Male	Female
A	125	100
B	210	165
C	276	212

3. Represent the following data by means of percentage sub-divided bar diagram.

Food crops	Area A (in 000,000 acres)	Area B (in 000,000 acres)
Rice	18	10
Wheat	12	14
Barley	10	8
Maize	7	6
Others	12	15

4. Draw a Pie diagram to exhibit the causes of death in the country.

Causes of Death	Numbers
Diarrhoea and enteritis	60
Prematurity and atrophy	170
Bronchitis and pneumonia	90

5. Draw a histogram and frequency polygon for the following data.

Weights (in kg)	Number of men
40-45	8
45-50	14
50-55	21
55-60	18
60-65	10

6. Draw a frequency curve for the following data.

Marks	No. of students
0-20	7
20-40	15
40-60	28
60-80	17
80-100	5

7. The frequency distribution of wages in a certain factory is as follows:

Wages	Number of workers
0- 500	10
500-1000	19
1000-1500	28
1500-2000	15
2000-2500	6

8. The following table given the weekly family income in two different region. Draw the Lorenz curve and compare the two regions of incomes.

Income	No. of families	
	Region A	Region B
1000	12	5
1250	18	10
1500	29	17
1750	42	23
2000	20	15
2500	11	8
3000	6	3

IV. Suggested Activities:

1. Give relevant diagrammatic representations for the activities listed in the previous lessons.
2. Get the previous monthly expenditure of your family and interpret it into bar diagram and pie diagram. Based on the data, propose a budget for the next month and interpreted into bar and pie diagram.

   Compare the two months expenditure through diagrams

Answers

I. 1. (a) 2. (a) 3.(d) 4. (d) 5.(b)

II.

1. Component bar
2. Area
3. Two
4. Median
5. Lorenz

6. MEASURES OF CENTRAL TENDENCY

Measures of Central Tendency:

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations. There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages.

The meaning of average is nicely given in the following definitions.

"A measure of central tendency is a typical value around which other figures congregate."

"An average stands for the whole group of which it forms a part yet represents the whole."

"One of the most widely used set of summary figures is known as measures of location."

Characteristics for a good or an ideal average :

The following properties should possess for an ideal average.

1. It should be rigidly defined.
2. It should be easy to understand and compute.
3. It should be based on all items in the data.
4. Its definition shall be in the form of a mathematical formula.
5. It should be capable of further algebraic treatment.
6. It should have sampling stability.
7. It should be capable of being used in further statistical computations or processing.

Besides the above requisites, a good average should represent maximum characteristics of the data, its value should be nearest to the most items of the given series.

Arithmetic mean or mean :

Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. If the variable x assumes n values x_1, x_2 …x_n then the mean, $\bar{x}$, is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

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

This formula is for the ungrouped or raw data.

Example 1 *:*

Calculate the mean for 2, 4, 6, 8, 10

Solution:

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5}$$

$$= \frac{30}{5} = 6$$

Short-Cut method :

Under this method an assumed or an arbitrary average (indicated by A) is used as the basis of calculation of deviations from individual values. The formula is

$$\bar{x} = A + \frac{\sum d}{n}$$

where, A = the assumed mean or any value in x

d = the deviation of each value from the assumed mean

Example 2 :

A student' s marks in 5 subjects are 75, 68, 80, 92, 56. Find his average mark.

Solution:

X	d=x-A
75	7
A 68	0
80	12
92	24
56	-12
Total	31

$$\bar{x} = A + \frac{\sum d}{n}$$

$$= 68 + \frac{31}{5}$$

$$= 68 + 6.2$$

$$= 74.2$$

Grouped Data :

The mean for grouped data is obtained from the following formula:

$$\bar{x} = \frac{\sum fx}{N}$$

where x = the mid-point of individual class

f = the frequency of individual class

N = the sum of the frequencies or total frequencies.

Short-cut method :

$$\bar{x} = A + \frac{\sum fd}{N} \times c$$

where $d = \frac{x - A}{c}$

A = any value in x

N = total frequency

c = width of the class interval

Example 3:

Given the following frequency distribution, calculate the arithmetic mean

Marks	: 64	63	62	61	60	59
Number of Students	: 8	18	12	9	7	6

Solution:

X	F	fx	d=x-A	fd
64	8	512	2	16
63	18	1134	1	18
62	12	744	0	0
61	9	549	–1	–9
60	7	420	–2	–14
59	6	354	–3	–18
	60	3713		- 7

Direct method

$$\bar{x} = \frac{\sum fx}{N} = \frac{3713}{60} = 61.88$$

Short-cut method

$$\bar{x} = A + \frac{\sum fd}{N} = 62 - \frac{7}{60} = 61.88$$

Example 4 *:*

Following is the distribution of persons according to different income groups. Calculate arithmetic mean.

Income Rs(100)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of persons	6	8	10	12	7	4	3

Solution:

Income C.I	Number of Persons (f)	Mid X	$d = \frac{x - A}{c}$	Fd
0-10	6	5	-3	-18
10-20	8	15	-2	-16
20-30	10	25	-1	-10
30-40	12	A 35	0	0
40-50	7	45	1	7
50-60	4	55	2	8
60-70	3	65	3	9
	50			-20

Mean = $\bar{x} = A + \frac{\sum fd}{N}$

$= 35 - \frac{20}{50} \times 10$

$= 35 - 4$

$= 31$

Merits and demerits of Arithmetic mean :

Merits:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits:

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

Weighted Arithmetic mean :

For calculating simple mean, we suppose that all the values or the sizes of items in the distribution have equal importance. But, in practical life this may not be so. In case some items are more

important than others, a simple average computed is not representative of the distribution. Proper weightage has to be given to the various items. For example, to have an idea of the change in cost of living of a certain group of persons, the simple average of the prices of the commodities consumed by them will not do because all the commodities are not equally important, e.g rice, wheat and pulses are more important than tea, confectionery etc., It is the weighted arithmetic average which helps in finding out the average value of the series after giving proper weight to each group.

Definition:

The average whose component items are being multiplied by certain values known as "weights" and the aggregate of the multiplied results are being divided by the total sum of their "weight".

If x_1, $x_2 . x_n$ be the values of a variable x with respective weights of w_1, $w_2 \ldots w_n$ assigned to them, then

Weighted A.M = $\bar{x}_w = \dfrac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n} = \dfrac{\sum w_i x_i}{\sum w_i}$

Uses of the weighted mean:

Weighted arithmetic mean is used in:

a. Construction of index numbers.
b. Comparison of results of two or more universities where number of students differ.
c. Computation of standardized death and birth rates.

Example 5:

Calculate weighted average from the following data

Designation	Monthly salary (in Rs)	Strength of the cadre
Class 1 officers	1500	10
Class 2 officers	800	20
Subordinate staff	500	70
Clerical staff	250	100
Lower staff	100	150

Solution:

Designation	Monthly salary,x	Strength of the cadre,w	wx
Class 1 officer	1,500	10	15,000
Class 2 officer	800	20	16,000
Subordinate staff	500	70	35,000
Clerical staff	250	100	25,000
Lower staff	100	150	15,000
		350	1,06,000

Weighted average, $\bar{x}_w = \frac{\sum wx}{\sum w}$

$= \frac{106000}{350}$

= Rs. 302.86

Harmonic mean (H.M) :

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If $x_1, x_2 \ldots x_n$ are n observations,

$$H.M = \frac{n}{\sum_{i=1}^{n} \left(\frac{1}{x_i}\right)}$$

For a frequency distribution

$$H.M. = \frac{N}{\sum_{i=1}^{n} f \left(\frac{1}{x_i}\right)}$$

Example 6:

From the given data calculate H.M 5,10,17,24,30

X	$\frac{1}{x}$
5	0.2000
10	0.1000
17	0.0588
24	0.0417
30	0.0333
Total	0.4338

$$\text{H.M} = \frac{n}{\sum\left[\frac{1}{x}\right]}$$

$$= \frac{5}{0.4338} = 11.526$$

Example 7*:*

The marks secured by some students of a class are given below. Calculate the harmonic mean.

Marks	20	21	22	23	24	25
Number of Students	4	2	7	1	3	1

Solution:

Marks X	No of students f	$\frac{1}{x}$	$f(\frac{1}{x})$
20	4	0.0500	0.2000
21	2	0.0476	0.0952
22	7	0.0454	0.3178
23	1	0.0435	0.0435
24	3	0.0417	0.1251
25	1	0.0400	0.0400
	18		0.8216

$$H.M = \frac{N}{\sum f\left[\frac{1}{x}\right]}$$

$$= \frac{18}{0.1968} = 21.91$$

Merits of H.M :

1. It is rigidly defined.
2. It is defined on all observations.
3. It is amenable to further algebraic treatment.
4. It is the most suitable average when it is desired to give greater weight to smaller observations and less weight to the larger ones.

Demerits of H.M :

1. It is not easily understood.
2. It is difficult to compute.
3. It is only a summary figure and may not be the actual item in the series
4. It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.

Geometric mean :

The geometric mean of a series containing n observations is the n^{th} root of the product of the values. If $x_1, x_2 \ldots x_n$ are observations then

$$G.M = \sqrt[n]{x_1 . x_2 \ldots x_n}$$

$$= (x_1 . x_2 \ldots x_n)^{1/n}$$

$$\log GM = \frac{1}{n} \log(x_1 . x_2 \ldots x_n)$$

$$= \frac{1}{n} (\log x_1 + \log x_2 + \ldots + \log x_n$$

$$= \frac{\sum \log x_i}{n}$$

$$GM = \text{Antilog} \frac{\sum \log x_i}{n}$$

For grouped data

$$GM = \text{Antilog}\left[\frac{\sum f \log x_i}{N}\right]$$

Example 8:

Calculate the geometric mean of the following series of monthly income of a batch of families 180,250,490,1400,1050

x	logx
180	2.2553
250	2.3979
490	2.6902
1400	3.1461
1050	3.0212
	13.5107

$$GM = \text{Antilog}\left[\frac{\sum \log x}{n}\right]$$

$$= \text{Antilog}\ \frac{13.5107}{5}$$

$$= \text{Antilog}\ 2.7021 \quad = 503.6$$

Example 9*:*

Calculate the average income per head from the data given below .Use geometric mean.

Class of people	Number of families	Monthly income per head (Rs)
Landlords	2	5000
Cultivators	100	400
Landless – labours	50	200
Money – lenders	4	3750
Office Assistants	6	3000
Shop keepers	8	750
Carpenters	6	600
Weavers	10	300

Solution:

Class of people	Annual income (Rs) X	Number of families (f)	Log x	f logx
Landlords	5000	2	3.6990	7.398
Cultivators	400	100	2.6021	260.210
Landless – labours	200	50	2.3010	115.050
Money – lenders	3750	4	3.5740	14.296
Office Assistants	3000	6	3.4771	20.863
Shop keepers	750	8	2.8751	23.2008
Carpenters	600	6	2.7782	16.669
Weavers	300	10	2.4771	24.771
		186		482.257

$$GM = \text{Antilog}\left[\frac{\sum f \log x}{N}\right]$$

$$= \text{Antilog}\left[\frac{482.257}{186}\right]$$

$$= \text{Antilog}\,(2.5928)$$

$$= \text{Rs } 391.50$$

Merits of Geometric mean :

1. It is rigidly defined
2. It is based on all items
3. It is very suitable for averaging ratios, rates and percentages
4. It is capable of further mathematical treatment.
5. Unlike AM, it is not affected much by the presence of extreme values

Demerits of Geometric mean:

1. It cannot be used when the values are negative or if any of the observations is zero
2. It is difficult to calculate particularly when the items are very large or when there is a frequency distribution.

3. It brings out the property of the ratio of the change and not the absolute difference of change as the case in arithmetic mean.
4. The GM may not be the actual value of the series.

Combined mean :

If the arithmetic averages and the number of items in two or more related groups are known, the combined or the composite mean of the entire group can be obtained by

$$\text{Combined mean } \bar{\bar{X}} = \left[\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \right]$$

The advantage of combined arithmetic mean is that, we can determine the over, all mean of the combined data without going back to the original data.

Example 10:

Find the combined mean for the data given below

$n_1 = 20$, $\bar{x}_1 = 4$, $n_2 = 30$, $\bar{x}_2 = 3$

Solution:

$$\text{Combined mean } \bar{\bar{X}} = \left[\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \right]$$

$$= \left[\frac{20 \times 4 + 30 \times 3}{20 + 30} \right]$$

$$= \left[\frac{80 + 90}{50} \right]$$

$$= \left[\frac{170}{50} \right] = 3.4$$

Positional Averages:

These averages are based on the position of the given observation in a series, arranged in an ascending or descending order. The magnitude or the size of the values does matter as was in the case of arithmetic mean. It is because of the basic difference

that the median and mode are called the positional measures of an average.

Median :

The median is that value of the variate which divides the group into two equal parts, one part comprising all values greater, and the other, all values less than median.

Ungrouped or Raw data :

Arrange the given values in the increasing or decreasing order. If the number of values are odd, median is the middle value .If the number of values are even, median is the mean of middle two values.

By formula

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

Example 11:

When odd number of values are given. Find median for the following data

25, 18, 27, 10, 8, 30, 42, 20, 53

Solution:

Arranging the data in the increasing order 8, 10, 18, 20, 25, 27, 30, 42, 53

The middle value is the 5^{th} item i.e., 25 is the median

Using formula

$$\text{Md} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{item.}$$

$$= \left(\frac{9+1}{2}\right)^{\text{th}} \text{item.}$$

$$= \left(\frac{10}{2}\right)^{\text{th}} \text{item}$$

$$= 5^{\text{th}} \text{ item}$$

$$= 25$$

Example 12 :

When even number of values are given. Find median for the following data

5, 8, 12, 30, 18, 10, 2, 22

Solution:

Arranging the data in the increasing order 2, 5, 8, 10, 12, 18, 22, 30

Here median is the mean of the middle two items (ie) mean of (10,12) ie

$$= \left(\frac{10+12}{2}\right) = 11$$

$\therefore$ median = 11.

Using the formula

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

$$= \left(\frac{8+1}{2}\right)^{\text{th}} \text{ item.}$$

$$= \left(\frac{9}{2}\right)^{\text{th}} \text{ item} = 4.5^{\text{th}} \text{ item}$$

$$= 4^{\text{th}} \text{ item} + \left(\frac{1}{2}\right)(5^{\text{th}} \text{ item} - 4^{\text{th}} \text{ item})$$

$$= 10 + \left(\frac{1}{2}\right)[12\text{-}10]$$

$$= 10 + \left(\frac{1}{2}\right) \times 2$$

$$= 10 + 1$$

$$= 11$$

Example 13*:*

The following table represents the marks obtained by a batch of 10 students in certain class tests in statistics and Accountancy.

Serial No	1	2	3	4	5	6	7	8	9	10

Marks (Statistics)	53	55	52	32	30	60	47	46	35	28
Marks (Accountancy)	57	45	24	31	25	84	43	80	32	72

Indicate in which subject is the level of knowledge higher ?

Solution:

For such question, median is the most suitable measure of central tendency. The mark in the two subjects are first arranged in increasing order as follows:

Serial No	1	2	3	4	5	6	7	8	9	10
Marks in Statistics	28	30	32	35	46	47	52	53	55	60
Marks in Accountancy	24	25	31	32	43	45	57	72	80	84

Median = $\left(\frac{n+1}{2}\right)^{th}$ item = $\left(\frac{10+1}{2}\right)^{th}$ item $= 5.5^{th}$ item

$$= \frac{Value\ of\ 5^{th}\ item + value\ of\ 6^{th}\ item}{2}$$

$$\text{Md (Statistics)} = \frac{46+47}{2} = 46.5$$

$$\text{Md (Accountancy)} = \frac{43+45}{2} = 44$$

There fore the level of knowledge in Statistics is higher than that in Accountancy.

Grouped Data:

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution , cumulative frequencies have to be calculated to know the total number of items.

Cumulative frequency : (cf)

Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, ie adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

Discrete Series:

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{N+1}{2}\right)$

Step3: See in the cumulative frequencies the value just greater than $\left(\frac{N+1}{2}\right)$

Step4: Then the corresponding value of x is median.

Example 14:

The following data pertaining to the number of members in a family. Find median size of the family.

Number of members **x**	1	2	3	4	5	6	7	8	9	10	11	12
Frequency F	1	3	5	6	10	13	9	5	3	2	2	1

Solution:

X	f	cf
1	1	1
2	3	4
3	5	9
4	6	15
5	10	25
6	13	38
7	9	47
8	5	52
9	3	55
10	2	57
11	2	59
12	1	60
	60	

Median = size of $\left(\frac{N+1}{2}\right)^{th}$ item

$= \text{size of}\left(\frac{60+1}{2}\right)^{th} \text{item}$

$= 30.5^{th}$ item

The cumulative frequencies just greater than 30.5 is 38.and the value of x corresponding to 38 is 6.Hence the median size is 6 members per family.

Note:

It is an appropriate method because a fractional value given by mean does not indicate the average number of members in a family.

Continuous Series:

The steps given below are followed for the calculation of median in continuous series.

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{N}{2}\right)$

Step3: See in the cumulative frequency the value first greater than $\left(\frac{N}{2}\right)$, Then the corresponding class interval is called the Median class. Then apply the formula

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

Where l = Lower limit of the median class

m = cumulative frequency preceding the median

c = width of the median class

f =frequency in the median class.

N=Total frequency.

Note :

If the class intervals are given in inclusive type convert them into exclusive type and call it as true class interval and consider lower limit in this.

Example 15:

The following table gives the frequency distribution of 325 workers of a factory, according to their average monthly income in a certain year.

Income group (in Rs)	Number of workers
Below 100	1
100-150	20
150-200	42
200-250	55
250-300	62
300-350	45
350-400	30
400-450	25
450-500	15
500-550	18
550-600	10
600 and above	2
	325

Calculate median income

Solution:

Income group (Class-interval)	Number of workers (Frequency)	Cumulative frequency c.f
Below 100	1	1
100-150	20	21
150-200	42	63
200-250	55	118
250-300	62	180
300-350	45	225
350-400	30	255
400-450	25	280
450-500	15	295
500-550	18	313
550-600	10	323
600 and above	2	325
	325	

$$\frac{N}{2} = \frac{325}{2} = 162.5$$

Here $l = 250$, $N = 325$, $f = 62$, $c = 50$, $m = 118$

$$\text{Md} = 250 + \left(\frac{162.5 - 118}{62} \right) \times 50$$

$$= 250 + 35.89$$

$$= 285.89$$

Example 16:

Calculate median from the following data

Value	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39
Frequency	5	8	10	12	7	6	3	2

Value	f	True class interval	c.f
0-4	5	0.5-4.5	5
5-9	8	4.5-9.5	13
10-14	10	9.5-14.5	23
15-19	12	14.5-19.5	35
20-24	7	19.5-24.5	42
25-29	6	24.5-29.5	48
30-34	3	29.5-34.5	51
35-39	2	34.5-39.5	53
	53		

$$\left(\frac{N}{2} \right) = \left(\frac{53}{2} \right) = 26.5$$

$$\text{Md} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 14.5 + \frac{26.5 - 23}{12} \times 5$$

$$= 14.5 + 1.46 = 15.96$$

Example 17:

Following are the daily wages of workers in a textile. Find the median.

Wages (in Rs.)	Number of workers
less than 100	5
less than 200	12
less than 300	20
less than 400	32
less than 500	40
less than 600	45
less than 700	52
less than 800	60
less than 900	68
less than 1000	75

Solution :

We are given upper limit and less than cumulative frequencies. First find the class-intervals and the frequencies. Since the values are increasing by 100, hence the width of the class interval equal to 100.

Class interval	f	c.f
0-100	5	5
100-200	7	12
200-300	8	20
300- 400	12	32
400-500	8	40
500-600	5	45
600-700	7	52
700-800	8	60
800-900	8	68
900-1000	7	75
	75	

$$\left(\frac{N}{2}\right)^{th} = \left(\frac{75}{2}\right) = 37.5$$

$$Md = l + \left(\frac{\frac{N}{2} - m}{f} \right) \times c$$

$$= 400 + \left(\frac{37.5 - 32}{8} \right) \times 100 = 400 + 68.75 = 468.75$$

Example 18:

Find median for the data given below.

Marks	Number of students
Greater than 10	70
Greater than 20	62
Greater than 30	50
Greater than 40	38
Greater than 50	30
Greater than 60	24
Greater than 70	17
Greater than 80	9
Greater than 90	4

Solution :

Here we are given lower limit and more than cumulative frequencies.

Class interval	f	More than c.f	Less than c.f
10-20	8	70	8
20-30	12	62	20
30-40	12	50	32
40-50	8	38	40
50-60	6	30	46
60-70	7	24	53
70-80	8	17	61
80-90	5	9	66
90-100	4	4	70
	70		

$$\left(\frac{N}{2} \right) = \left(\frac{70}{2} \right) = 35$$

$$\text{Median} = l + \frac{\left(\frac{N}{2} - m\right)}{f} \times c$$

$$= 40 + \frac{(35 - 32)}{8} \times 10$$

$$= 40 + 3.75$$

$$= 43.75$$

Example 19:
Compute median for the following data.

Mid-Value	5	15	25	35	45	55	65	75
Frequency	7	10	15	17	8	4	6	7

Solution :
Here values in multiples of 10, so width of the class interval is 10.

Mid x	C.I	f	c.f
5	0-10	7	7
15	10-20	10	17
25	20-30	15	32
35	30-40	17	49
45	40-50	8	57
55	50-60	4	61
65	60-70	6	67
75	70-80	7	74
		74	

$$\left(\frac{N}{2}\right) = \left(\frac{74}{2}\right) = 37$$

$$\text{Median} = l + \frac{\left(\frac{N}{2} - m\right)}{f} \times c$$

$$= 30 + \left(\frac{37 - 32}{17}\right) \times 10$$

$$= 30 + 2.94$$

$$= 32.94$$

Graphic method for Location of median:

Median can be located with the help of the cumulative frequency curve or ‘ ogive’ . The procedure for locating median in a grouped data is as follows:

Step1: The class boundaries, where there are no gaps between consecutive classes, are represented on the horizontal axis (x-axis).

Step2: The cumulative frequency corresponding to different classes is plotted on the vertical axis (y-axis) against the upper limit of the class interval (or against the variate value in the case of a discrete series.)

Step3: The curve obtained on joining the points by means of freehand drawing is called the ‘ ogive’ . The ogive so drawn may be either a (i) less than ogive or a (ii) more than ogive.

Step4: The value of $\frac{N}{2}$ or $\frac{N+1}{2}$ is marked on the y-axis, where N is the total frequency.

Step5: A horizontal straight line is drawn from the point $\frac{N}{2}$ or $\frac{N+1}{2}$ on the y-axis parallel to x-axis to meet the ogive.

Step6: A vertical straight line is drawn from the point of intersection perpendicular to the horizontal axis.

Step7: The point of intersection of the perpendicular to the x-axis gives the value of the median.

Remarks :

1. From the point of intersection of ‘ less than’ and ‘ more than’ ogives, if a perpendicular is drawn on the x-axis, the point so obtained on the horizontal axis gives the value of the median.
2. If ogive is drawn using cumulated percentage frequencies, then we draw a straight line from the point intersecting 50

percent cumulated frequency on the y-axis parallel to the x-axis to intersect the ogive. A perpendicular drawn from this point of intersection on the horizontal axis gives the value of the median.

Example 20:

Draw an ogive of ' less than' type on the data given below and hence find median.

Weight(lbs)	Number of persons
100-109	8
110-119	15
120-129	21
130-139	34
140-149	45
150-159	26
160-169	20
170-179	15
180-189	10
190-199	6

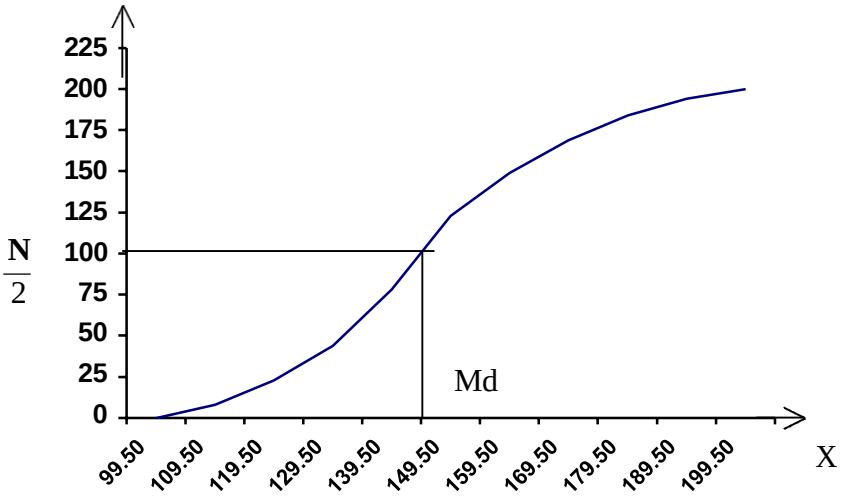
Solution:

Class interval	No of persons	True class interval	Less than c.f
100-109	8	99.5-109.5	8
110-119	15	109.5-119.5	23
120-129	21	119.5-129.5	44
130-139	34	129.5-139.5	78
140-149	45	139.5-149.5	123
150-159	26	149.5-159.5	149
160-169	20	159.5-169.5	169
170-179	15	169.5-179.5	184
180-189	10	179.5-189.5	194
190-199	6	189.5-199.5	200

Y

Less than Ogive

X axis 1cm = 10 units
Y axis 1cm = 25 units



Example 21:

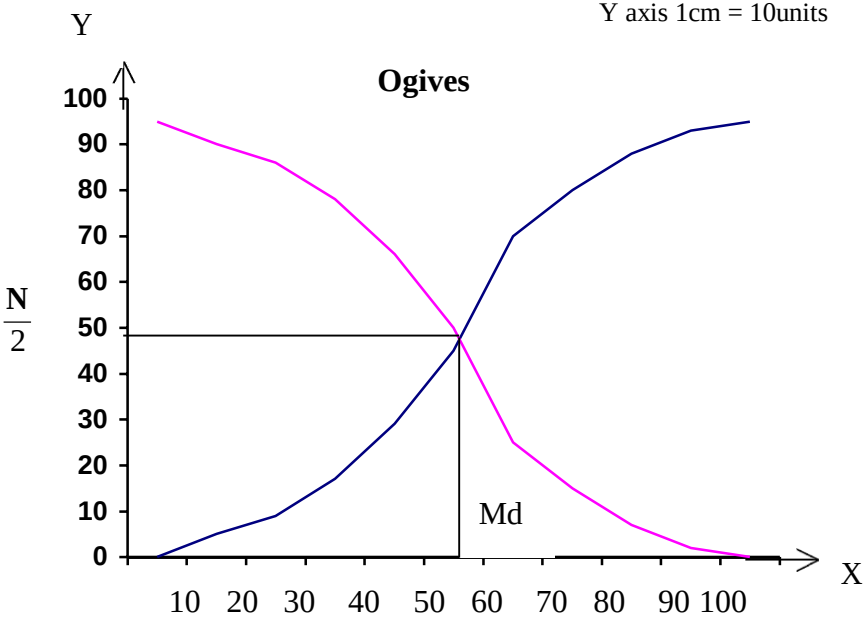
Draw an ogive for the following frequency distribution and hence find median.

Marks	Number of students
0-10	5
10-20	4
20-30	8
30-40	12
40-50	16
50-60	25
60-70	10
70-80	8
80-90	5
90-100	2

Solution:

Class boundary	Cumulative Frequency	
	Less than	More than
0	0	95
10	5	90
20	9	86
30	17	78
40	29	66
50	45	50
60	70	25
70	80	15
80	88	7
90	93	2
100	95	0

X axis 1cm = 10units
Y axis 1cm = 10units



Merits of Median :

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open-end intervals.
3. Median can be located even if the data are incomplete.
4. Median can be located even for qualitative factors such as ability, honesty etc.

Demerits of Median :

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in mean deviation.
4. It is not taken into account all the observations.

Quartiles :

The quartiles divide the distribution in four parts. There are three quartiles. The second quartile divides the distribution into two halves and therefore is the same as the median. The first (lower) quartile (Q_1) marks off the first one-fourth, the third (upper) quartile (Q_3) marks off the three-fourth.

Raw or ungrouped data:

First arrange the given data in the increasing order and use the formula for Q_1 and Q_3 then quartile deviation, Q.D is given by

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

Where $Q_1 = \left(\frac{n+1}{4}\right)^{th}$ item and $Q_3 = 3\left(\frac{n+1}{4}\right)^{th}$ item

Example 22 :

Compute quartiles for the data given below 25,18,30, 8, 15, 5, 10, 35, 40, 45

Solution :

5, 8, 10, 15, 18,25, 30,35,40, 45

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

$$= \left(\frac{10+1}{4}\right)^{th} \text{item}$$

$$= (2.75)^{th} \text{item}$$

$$= 2^{nd} \text{item} + \left(\frac{3}{4}\right) (3^{rd} \text{item} - 2^{nd} \text{item})$$

$$= 8 + \frac{3}{4} (10-8)$$

$$= 8 + \frac{3}{4} \times 2$$

$$= 8 + 1.5$$

$$= 9.5$$

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

$$= 3 \times (2.75)^{th} \text{item}$$

$$= (8.25)^{th} \text{item}$$

$$= 8^{th} \text{item} + \frac{1}{4} [9^{th} \text{item} - 8^{th} \text{item}]$$

$$= 35 + \frac{1}{4} [40-35]$$

$$= 35+1.25=36.25$$

Discrete Series :

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{N+1}{4}\right)$

Step3: See in the cumulative frequencies , the value just greater than $\left(\frac{N+1}{4}\right)$,then the corresponding value of x is Q_1

Step4: Find $3\left(\frac{N+1}{4}\right)$

Step5: See in the cumulative frequencies, the value just greater than $3\left(\frac{N+1}{4}\right)$, then the corresponding value of x is Q_3

Example 23:

Compute quartiles for the data given bellow.

X	5	8	12	15	19	24	30
f	4	3	2	4	5	2	4

Solution:

x	f	c.f ←
5	4	4
8	3	7
12	2	9
15	4	13
19	5	18
24	2	20
30	4	24
Total	24	

$Q_1 = \left(\frac{N+1}{4}\right)^{th}$ item $= \left(\frac{24+1}{4}\right) = \left(\frac{25}{4}\right) = 6.25^{th}$ item

$Q_3 = 3\left(\frac{N+1}{4}\right)^{th}$ item $= 3\left(\frac{24+1}{4}\right) = 18.75^{th}$ item $\therefore Q_1 = 8$; $Q_3 = 24$

Continuous series :

Step1: Find cumulative frequencies

Step2: Find $\left(\frac{N}{4}\right)$

Step3: See in the cumulative frequencies, the value just greater than $\left(\frac{N}{4}\right)$, then the corresponding class interval is called first quartile class.

Step4: Find $3\left(\frac{N}{4}\right)$ See in the cumulative frequencies the value just greater than $3\left(\frac{N}{4}\right)$ then the corresponding class interval is called 3rd quartile class. Then apply the respective formulae

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$$Q_3 = l_3 + \frac{3\left(\frac{N}{4}\right) - m_3}{f_3} \times c_3$$

Where l_1 = lower limit of the first quartile class
f_1 = frequency of the first quartile class
c_1 = width of the first quartile class
m_1 = c.f. preceding the first quartile class
l_3 = 1ower limit of the 3rd quartile class
f_3 = frequency of the 3rd quartile class
c_3 = width of the 3rd quartile class
m_3 = c.f. preceding the 3rd quartile class

Example 24:

The following series relates to the marks secured by students in an examination.

Marks	No. of students
0-10	11
10-20	18
20-30	25
30-40	28
40-50	30
50-60	33
60-70	22
70-80	15
80-90	12
90-100	10

Find the quartiles

Solution :

C.I.	f	cf
0-10	11	11
10-20	18	29
20-30	25	54
30-40	28	82
40-50	30	112
50-60	33	145
60-70	22	167
70-80	15	182
80-90	12	194
90-100	10	204
	204	

$\left(\frac{N}{4}\right) = \left(\frac{204}{4}\right) = 51 \qquad 3\left(\frac{N}{4}\right) = 153$

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$$= 20 + \frac{51 - 29}{25} \times 10 \qquad = 20 + 8.8 = 28.8$$

$$Q_3 = l_3 + \frac{3\left(\frac{N}{4}\right) - m_3}{f_3} \times c_3$$

$$= 60 + \frac{153 - 145}{22} \times 12 = 60 + 4.36 = 64.36$$

Deciles :

These are the values, which divide the total number of observation into 10 equal parts. These are 9 deciles D_1, $D_2 \ldots D_9$. These are all called first decile, second decile. etc.,

Deciles for Raw data or ungrouped data

Example 25:

Compute D_5 for the data given below

5, 24, 36, 12, 20, 8

Solution :

Arranging the given values in the increasing order

5, 8, 12, 20, 24, 36

$$D_5 = \left(\frac{5(n+1)}{10}\right)^{th} \text{observation}$$

$$= \left(\frac{5(6+1)}{10}\right)^{th} \text{observation}$$

$$= (3.5)^{th} \text{ observation}$$

$$= 3^{rd} \text{ item} + \frac{1}{2}\,[4^{th} \text{ item} - 3^{rd} \text{ item}]$$

$$= 12 + \frac{1}{2}\,[20 - 12] \quad = 12 + 4 = 16$$

Deciles for Grouped data :

Example 26:

Calculate D_3 and D_7 for the data given below

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency :	5	7	12	16	10	8	4

Solution :

C.I	f	c.f
0-10	5	5
10-20	7	12
20-30	12	24
30-40	16	40
40-50	10	50
50-60	8	58
60-70	4	62
	62	

D_3 item $= \left(\frac{3N}{10}\right)^{th}$ item

$= \left(\frac{3\times 62}{10}\right)^{th}$ item

$= (18.6)^{th}$ item

which lies in the interval 20-30

$$\therefore D_3 = l + \frac{3\left(\frac{N}{10}\right) - m}{f} \times c$$

$$= 20 + \frac{18.6 - 12}{12} \times 10$$

$$= 20 + 5.5 = 25.5$$

D_7 item $= \left(\frac{7 \times N}{10}\right)^{th}$ item

$= \left(\frac{7 \times 62}{10}\right)^{th}$ item

$= \left(\frac{434}{10}\right)^{th}$ item $= (43.4)^{th}$ item

which lies in the interval(40-50)

$$D_7 = l + \frac{\left(\frac{7N}{10}\right) - m}{f} \times c$$

$$= 40 + \frac{43.4 - 40}{10} \times 10$$

$$= 40 + 3.4 = 43.4$$

Percentiles :

The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The percentile (P_k) is that value of the variable up to which lie exactly k% of the total number of observations.

Relationship :

$P_{25} = Q_1$; $P_{50} = D_5 = Q_2$ = Median and $P_{75} = Q_3$

Percentile for Raw Data or Ungrouped Data :

Example 27:

Calculate P_{15} for the data given below:

5, 24 , 36 , 12 , 20 , 8

Arranging the given values in the increasing order.

5, 8, 12, 20, 24, 36

$$P_{15} = \left(\frac{15(n+1)}{100} \right)^{th} \text{item}$$

$$= \left(\frac{15 \times 7}{100} \right)^{th} \text{item}$$

$= (1.05)^{th}$ item

$= 1^{st}$ item + 0.05 (2^{nd} item – 1^{st} item)

$= 5 + 0.05\ (8-5)$

$= 5 + 0.15 = 5.15$

Percentile for grouped data :

Example 28:

Find P_{53} for the following frequency distribution.

Class interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	5	8	12	16	20	10	4	3

Solution:

Class Interval	Frequency	C.f
0-5	5	5
5-10	8	13
10-15	12	25
15-20	16	41
20-25	20	61
25-30	10	71
30-35	4	75
35-40	3	78
Total	78	

$$P_{53} = l + \frac{\frac{53N}{100} - m}{f} \times c$$

$$= 20 + \frac{41.34 - 41}{20} \times 5$$

$$= 20 + 0.085 = 20.085.$$

Mode :

The mode refers to that value in a distribution, which occur most frequently. It is an actual value, which has the highest concentration of items in and around it.

According to Croxton and Cowden " The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded at the most typical of a series of values".

It shows the centre of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Its importance is very great in marketing studies where a manager is interested in knowing about the size, which has the highest concentration of items. For example, in placing an order for shoes or ready-made garments the modal size helps because this sizes and other sizes around in common demand.

Computation of the mode:

Ungrouped or Raw Data:

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example 29*:*

2 , 7, 10, 15, 10, 17, 8, 10, 2

$\therefore$ Mode = $M_0 = 10$

In some cases the mode may be absent while in some cases there may be more than one mode.

Example 30:

1. 12, 10, 15, 24, 30 (no mode)
2. 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

$\therefore$ the modes are 7 and 10

Grouped Data:

For Discrete distribution, see the highest frequency and corresponding value of X is mode.

Continuous distribution :

See the highest frequency then the corresponding value of class interval is called the modal class. Then apply the formula.

$$\text{Mode} = M_0 = l + \frac{\triangle_1}{\triangle_1 + \triangle_2} \times C$$

l = Lower limit of the model class

$$\triangle_1 = f_1 - f_0$$
$$\triangle_2 = f_1 - f_2$$

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

The above formula can also be written as

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

Remarks :

1. If $(2f_1 - f_0 - f_2)$ comes out to be zero, then mode is obtained by the following formula taking absolute differences within vertical lines.
2. $M_0 = l + \frac{(f_1 - f_0)}{|f_1 - f_0| + |f_1 - f_2|} \times c$
3. If mode lies in the first class interval, then f_0 is taken as zero.

4. The computation of mode poses no problem in distributions with open-end classes, unless the modal value lies in the open-end class.

Example 31:

Calculate mode for the following :

C- I	f
0-50	5
50-100	14
100-150	40
150-200	91
200-250	150
250-300	87
300-350	60
350-400	38
400 and above	15

Solution:

The highest frequency is 150 and corresponding class interval is 200 – 250, which is the modal class.

Here l=200,f_1=150,f_0=91, f_2=87, C=50

$$\text{Mode} = M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

$$= 200 + \frac{150-91}{2\times150 - 91 - 87} \times 50$$

$$= 200 + \frac{2950}{122}$$

$$= 200 + 24.18 = 224.18$$

Determination of Modal class :

For a frequency distribution modal class corresponds to the maximum frequency. But in any one (or more) of the following cases

i. If the maximum frequency is repeated
ii. If the maximum frequency occurs in the beginning or at the end of the distribution
iii. If there are irregularities in the distribution, the modal class is determined by the method of grouping.

Steps for Calculation :

We prepare a grouping table with 6 columns

1. In column I, we write down the given frequencies.
2. Column II is obtained by combining the frequencies two by two.
3. Leave the 1st frequency and combine the remaining frequencies two by two and write in column III
4. Column IV is obtained by combining the frequencies three by three.
5. Leave the 1st frequency and combine the remaining frequencies three by three and write in column V
6. Leave the 1st and 2nd frequencies and combine the remaining frequencies three by three and write in column VI

Mark the highest frequency in each column. Then form an analysis table to find the modal class. After finding the modal class use the formula to calculate the modal value.

Example 32:

Calculate mode for the following frequency distribution.

Class interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	9	12	15	16	17	15	10	13

Grouping Table

C I	f	2	3	4	5	6
0- 5	9					
5-10	12	21				
10-15	15		27	36		
15-20	16	31			**43**	
20-25	**17**		**33**			**48**
25-30	15	**32**		**48**		
30-35	10		25		42	38
35-40	13	23				

Analysis Table

Columns	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
1					1			
2					1	1		
3				1	1			
4				1	1	1		
5		1	1	1				
6			1	1	1			
Total		1	2	4	5	2		

The maximum occurred corresponding to 20-25, and hence it is the modal class.

$$\text{Mode} = \text{Mo} = l + \frac{\triangle_1}{\triangle_1 + \triangle_2} \times C$$

Here $l = 20$; $\triangle_1 = f_1 - f_0 = 17 - 16 = 1$

$\triangle_2 = f_1 - f_2 = 17 - 15 = 2$

$$\therefore M_0 = 20 + \frac{1}{1+2} \times 5$$

$$= 20 + 1.67 = 21.67$$

Graphic Location of mode:

Steps:

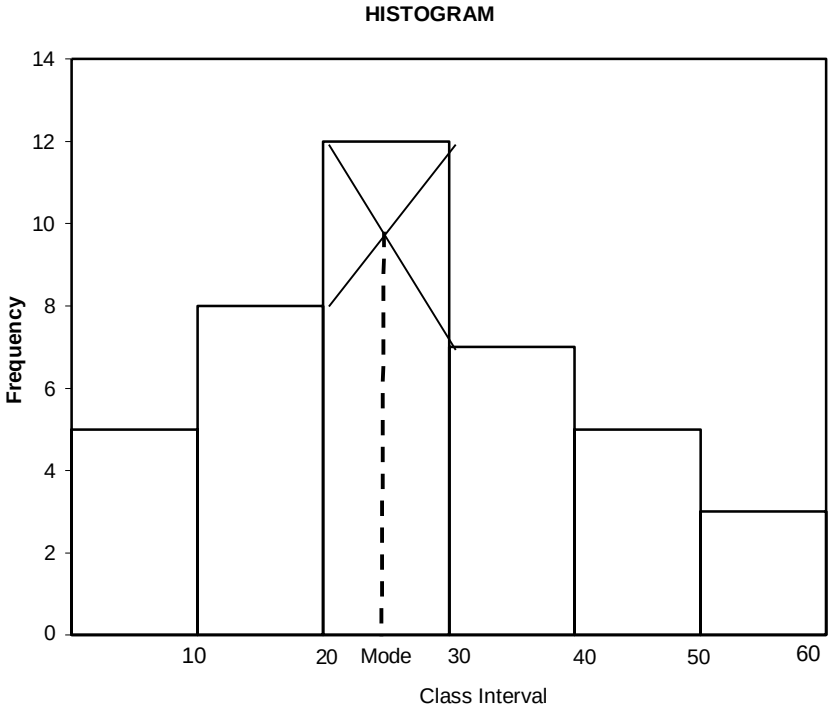
1. Draw a histogram of the given distribution.
2. Join the rectangle corner of the highest rectangle (modal class rectangle) by a straight line to the top right corner of the preceding rectangle. Similarly the top left corner of the highest rectangle is joined to the top left corner of the rectangle on the right.
3. From the point of intersection of these two diagonallines, draw a perpendicular to the x -axis.
4. Read the value in x-axis gives the mode.

Example 33:

Locate the modal value graphically for the following frequency distribution.

Class interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	5	8	12	7	5	3

Solution:



Merits of Mode:

1. It is easy to calculate and in some cases it can be located mere inspection
2. Mode is not at all affected by extreme values.
3. It can be calculated for open-end classes.
4. It is usually an actual value of an important part of the series.
5. In some circumstances it is the best representative of data.

Demerits of mode:

1. It is not based on all observations.
2. It is not capable of further mathematical treatment.
3. Mode is ill-defined generally, it is not possible to find mode in some cases.

4. As compared with mean, mode is affected to a great extent, by sampling fluctuations.
5. It is unsuitable in cases where relative importance of items has to be considered.

EMPIRICAL RELATIONSHIP BETWEEN AVERAGES

In a symmetrical distribution the three simple averages mean = median = mode. For a moderately asymmetrical distribution, the relationship between them are brought by Prof. Karl Pearson as mode = 3median - 2mean.

Example 34:

If the mean and median of a moderately asymmetrical series are 26.8 and 27.9 respectively, what would be its most probable mode?

Solution:

Using the empirical formula

Mode = 3 median – 2 mean

$= 3 \times 27.9 - 2 \times 26.8$

$= 30.1$

Example 35:

In a moderately asymmetrical distribution the values of mode and mean are 32.1 and 35.4 respectively. Find the median value.

Solution:

Using empirical Formula

$$\text{Median} = \frac{1}{3}[2\text{mean}+\text{mode}]$$

$$= \frac{1}{3}[2 \times 35.4 + 32.1]$$

$$= 34.3$$

Exercise - 6

I Choose the correct answer:

1. Which of the following represents median?
   a) First Quartile b) Fiftieth Percentile
   c) Sixth decile d) Third quartile

2. If the grouped data has open-end classes, one can not calculate.
   a)median b) mode c) mean d) quartile
3. Geometric mean of two numbers $\left(\frac{1}{16}\right)$ and $\left(\frac{4}{25}\right)$ is
   a) $\left(\frac{1}{10}\right)$ b) $\left(\frac{1}{100}\right)$ c) 10 d) 100
4. In a symmetric distribution
   a) mean ≠ median ≠ mode b) mean = median = mode
   c) mean > median > mode d) mean< median < mode
5. If modal value is not clear in a distribution , it can be ascertained by the method of
   a) grouping b) guessing
   c) summarizing d) trial and error
6. Shoe size of most of the people in India is No. 7 . Which measure of central value does it represent ?
   a).mean b) second quartile
   c) eighth decile d) mode
7. The middle value of an ordered series is called :
   a). 2^{nd} quartile b) 5^{th} decile
   c) 50^{th} percentile d) all the above
8. The variate values which divide a series (frequency distribution) into ten equal parts are called :
   a). quartiles b) deciles c) octiles d) percentiles
9. For percentiles, the total number of partition values are
   a) 10 b) 59 c) 100 d) 99
10. The first quartile divides a frequency distribution in the ratio
    a) 4 : 1 b) 1 :4 c) 3 : 1 d) 1 : 3
11. Sum of the deviations about mean is
    a) Zero b) minimum c) maximum d) one
12. Histogram is useful to determine graphically the value of
    a) mean b) median c)mode d)all the above
13. Median can be located graphically with the help of
    a) Histogram b) ogives
    c) bar diagram d) scatter diagram

14. Sixth deciles is same as
    a) median b) 50^{th} percentile
    c) 60^{th} percentile d) first quartile
15. What percentage of values lies between 5^{th} and 25^{th} percentiles?
    a) 5% b) 20%
    c) 30% d) 75%

II Fill in the blanks:

16. If 5 is subtracted from each observation of a set, then the mean of the observation is reduced by _____
17. The arithmetic mean of ***n*** natural numbers from 1 to *n* is____
18. Geometric mean cannot be calculated if any value of the set is _____
19. Median is a more suited average for grouped data with ____ classes.
20. 3^{rd} quartile and_____percentile are the same.

III Answer the following questions:

21. What do you understand by measures of central tendency?
22. What are the desirable characteristics of a good measure of central tendency.
23. What is the object of an average?
24. Give two examples where (i)Geometric mean and(ii)Harmonic mean would be most suitable averages.
25. Define median .Discuss its advantages and disadvantages as an average.
26. The monthly income of ten families(in rupees) in a certain locality are given below.

Family	A	B	C	D	E	F	G
Income(in rupees)	30	70	60	100	200	150	300

Calculate the arithmetic average by
(a)Direct method and (b)Short-cut method

27. Calculate the mean for the data

X:	5	8	12	15	20	24
f:	3	4	6	5	3	2

28. The following table gives the distribution of the number of workers according to the weekly wage in a company.

Weekly wage (in Rs.100's)	0-10	10-20	20-30	30-40
Numbers of workers	5	10	15	18

40-50	50-60	60-70	70-80
7	8	5	3

Obtain the mean weekly wage.

29. Mean of 20 values is 45. If one of these values is to be taken 64 instead of 46, find the corrected mean (ans:44.1)

30. From the following data, find the missing frequency when mean is 15.38

Size :	10	12	14	16	18	20
Frequency:	3	7	___	20	8	5

31. The following table gives the weekly wages in rupees of workers in a certain commercial organization. The frequency of the class-interval 49-52 is missing.

Weekly wages (in rs) :	40-43	43-46	46-49	49-52	52-55
Number of workers	31	58	60	___	27

It is known that the mean of the above frequency distribution is Rs .47.2. Find the missing frequency.

32. Find combined mean from the following data

$\overline{X}_1 = 210$ $n_1=50$

$X_2 = 150$ $n_2 =100$

33. Find combined mean from the following data

Group	1	2	3
Number	200	250	300
Mean	25	10	15

34. Average monthly production of a certain factory for the first 9 months is 2584, and for remaining three months it is 2416 units. Calculate average monthly production for the year.

35. The marks of a student in written and oral tests in subjects A, B and C are as follows. The written test marks are out of 75 and the oral test marks are out of 25. Find the weighted mean of the marks in written test taking the marks in oral test as weight. The marks of written test and oral test respectively as follows: 27, 24, 43 and 5, 10, 15.

36. The monthly income of 8 families is given below. Find GM.

Family :	A	B	C	D	E	F	G	H
Income(Rs)	70	10	500	75	8	250	8	42

37. The following table gives the diameters of screws obtained in a sample inquiry. Calculate the mean diameter using geometric average.

Diameter(m.m)	130	135	140	145	146	148	149	150	157
No. of. Screws	3	4	6	6	3	5	2	1	1

38. An investor buys Rs.1, 200 worth of shares in a company each month. During the first 5 months he bought the shares at a price of Rs.10, Rs.12, Rs.15, Rs.20 and Rs.24 per share. After 5 months what is the average price paid for the shares by him.

39. Determine median from the following data

25, 20, 15, 45, 18, 7, 10, 38, 12

40. Find median of the following data

Wages (in Rs)	60-70	50-60	40-50	30-40	20-30
Number of workers	7	21	11	6	5

41. The table below gives the relative frequency distribution of annual pay roll for 100 small retail establishments in a city.

Annual pay roll (1000 rupees)	Establishments
Less than 10	8
10 and Less than 20	12
20 and Less than 30	18
30 and Less than 40	30
40 and Less than 50	20
50 and Less than 60	12
	100

Calculate Median pay.

42. Calculate the median from the data given below

Wages (in Rs)	Number of workers	Wages (in Rs)	Number of workers
Above 30	520	Above 70	105
Above 40	470	Above 80	45
Above 50	399	Above 90	7
Above 60	210		

43. From the following data, compute the values of upper and lower quartiles, median, D_6, P_{20}.

Marks	No. of. Students	Marks	No. of. Students
Below 10	5	40-50	90
10-20	25	50-60	40
20-30	40	60-70	20
30-40	70	Above 70	10

44. Draw an ogive curve from the following data to find out the values of median and upper and lower quartiles.

Classes	90-100	100-110	110-120	120-130	130-140	140-150	150-160
Frequency	16	22	45	60	50	24	10

45. Calculate mode from the following data

Income (Rs)	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of. Persons	24	42	56	66	108	130	154

46. Represent the following data by means of histogram and from it, obtain value of mode.

Weekly wages (Rs)	10-15	15-20	20-25	25-30	30-35	35-40	40-45
No. of. Workers	7	9	27	15	12	12	8

Suggested Activities:

1. Measure the heights and weights of your class students. Find the mean, median, mode and compare
2. Find the mean marks of your class students in various subjects.

Answers:

I

1. (b) 2. (c) 3. (a) 4. (b)
5. (a) 6. (d) 7. (d) 8. (b)
9. (c) 10. (d) 11. (a) 12. (c)
13. (b) 14. (c) 15. (b)

II

16. 5 17 $\left(\frac{n+1}{2}\right)$ 18. 0 and negative
19. Open end 20 75^{th}

III

26. 130 27. 13.13 28. 35 29. 44.1
30. 12 31. 44 32. 170 33. 16
34 2542 35. 34 36. G.M.= 45.27
37. 142.5 mm 38. Rs.14.63 39. MD= 18 40.51.42
41. 34 42. 57.3
43. Q_1=30.714;Q_2=49.44; MD=41.11;D_6=44.44;P_{20}=27.5
44. MD=125.08; Q_1=114.18; Q_3=135.45
45 Mode=71.34

7. MEASURES OF DISPERSION – SKEWNESS AND KURTOSIS

Introduction :

The measure of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity. For example consider the following marks of two students.

Student I	Student II
68	85
75	90
65	80
67	25
70	65

Both have got a total of 345 and an average of 69 each. The fact is that the second student has failed in one paper. When the averages alone are considered, the two students are equal. But first student has less variation than second student. Less variation is a desirable characteristic.

Characteristics of a good measure of dispersion:

An ideal measure of dispersion is expected to possess the following properties

1. It should be rigidly defined
2. It should be based on all the items.
3. It should not be unduly affected by extreme items.

4. It should lend itself for algebraic manipulation.
5. It should be simple to understand and easy to calculate

Absolute and Relative Measures :

There are two kinds of measures of dispersion, namely

1.Absolute measure of dispersion

2.Relative measure of dispersion.

Absolute measure of dispersion indicates the amount of variation in a set of values in terms of units of observations. For example, when rainfalls on different days are available in mm, any absolute measure of dispersion gives the variation in rainfall in mm. On the other hand relative measures of dispersion are free from the units of measurements of the observations. They are pure numbers. They are used to compare the variation in two or more sets, which are having different units of measurements of observations.

The various absolute and relative measures of dispersion are listed below.

Absolute measure	**Relative measure**
1. Range	1.Co-efficient of Range
2. Quartile deviation	2.Co-efficient of Quartile deviation
3.Mean deviation	3. Co-efficient of Mean deviation
4.Standard deviation	4.Co-efficient of variation

Range and coefficient of Range:

Range:

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols, Range = L – S.

Where L = Largest value.

S = Smallest value.

In individual observations and discrete series, L and S are easily identified. In continuous series, the following two methods are followed.

Method 1:

L = Upper boundary of the highest class

S = Lower boundary of the lowest class.

Method 2:

L = Mid value of the highest class.

S = Mid value of the lowest class.

Co-efficient of Range :

$$\text{Co-efficient of Range} = \frac{L - S}{L + S}$$

Example1:

Find the value of range and its co-efficient for the following data.

7, 9, 6, 8, 11, 10, 4

Solution:

L=11, S = 4.

Range = L – S = 11- 4 = 7

$$\text{Co-efficient of Range} = \frac{L - S}{L + S}$$

$$= \frac{11 - 4}{11 + 4}$$

$$= \frac{7}{15} = 0.4667$$

Example 2:

Calculate range and its co efficient from the following distribution.

Size:	60-63	63-66	66-69	69-72	72-75
Number:	5	18	42	27	8

Solution:

L = Upper boundary of the highest class.

= 75

S = Lower boundary of the lowest class.
= 60

Range = L – S = 75 – 60 = 15

$$\text{Co-efficient of Range} = \frac{L - S}{L + S}$$

$$= \frac{75 - 60}{75 + 60}$$

$$= \frac{15}{135} = 0.1111$$

Merits and Demerits of Range :

Merits:

1. It is simple to understand.
2. It is easy to calculate.
3. In certain types of problems like quality control, weather forecasts, share price analysis, et c., range is most widely used.

Demerits:

1. It is very much affected by the extreme items.
2. It is based on only two extreme observations.
3. It cannot be calculated from open-end class intervals.
4. It is not suitable for mathematical treatment.
5. It is a very rarely used measure.

Quartile Deviation and Co efficient of Quartile Deviation :

Quartile Deviation (Q.D) :

Definition: Quartile Deviation is half of the difference between the first and third quartiles. Hence, it is called Semi Inter Quartile Range.

In Symbols, $Q.D = \frac{Q_3 - Q_1}{2}$. Among the quartiles Q_1, Q_2 and Q_3, the range $Q_3 - Q_1$ is called inter quartile range and $\frac{Q_3 - Q_1}{2}$, Semi inter quartile range.

Co-efficient of Quartile Deviation :

$$\text{Co-efficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example 3:

Find the Quartile Deviation for the following data:
391, 384, 591, 407, 672, 522, 777, 733, 1490, 2488

Solution:

Arrange the given values in ascending order.

384, 391, 407, 522, 591, 672, 733, 777, 1490, 2488.

Position of Q_1 is $\frac{n+1}{4} = \frac{10+1}{4} = 2.75^{th}$ item

$Q_1 = 2^{nd}$ value + 0.75 (3^{rd} value – 2^{nd} value)
$= 391 + 0.75\ (407 - 391)$
$= 391 + 0.75 \times 16$
$= 391 + 12$
$= 403$

Position Q_3 is $3\ \frac{n+1}{4} = 3 \times 2.75 = 8.25^{th}$ item

$Q_3 = 8^{th}$ value + 0.25 (9^{th} value – 8^{th} value)
$= 777 + 0.25\ (1490 - 777)$
$= 777 + 0.25\ (713)$
$= 777 + 178.25 = 955.25$

$$\text{Q.D} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{955.25 - 403}{2}$$

$$= \frac{552.25}{2} = 276.125$$

Example 4 :

Weekly wages of labours are given below. Calculated Q.D and Coefficient of Q.D.

Weekly Wage (Rs.)	:100	200	400	500	600
No. of Weeks	: 5	8	21	12	6

Solution :

Weekly Wage (Rs.)	No. of Weeks	Cum. No. of Weeks
100	5	5
200	8	13
400	21	34
500	12	46
600	6	52
Total	N=52	

Position of Q_1 in $\frac{N+1}{4} = \frac{52+1}{4} = 13.25^{th}$ item

Q_1 $= 13^{th}$ value + 0.25 (14^{th} Value – 13^{th} value)
$= 13^{th}$ value + 0.25 (400 – 200)
$= 200 + 0.25\ (400 - 200)$
$= 200 + 0.25\ (200)$
$= 200 + 50 = 250$

Position of Q is $3\left(\frac{N+1}{4}\right) = 3 \times 13.25 = 39.75^{th}$ item

Q_3 $= 39^{th}$ value + 0.75 (40^{th} value – 39^{th} value)
$= 500 + 0.75\ (500 - 500)$
$= 500 + 0.75 \times 0$
$= 500$

Q.D. $= \frac{Q_3 - Q_1}{2} = \frac{500 - 250}{2} = \frac{250}{2} = 125$

Coefficient of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

$= \frac{500 - 250}{500 + 250}$

$= \frac{250}{750} = 0.3333$

Example 5:

For the date given below, give the quartile deviation and coefficient of quartile deviation.

X :	351 – 500	501 – 650	651 – 800	801–950	951–1100
f :	48	189	88	4	28

Solution :

x	f	True class Intervals	Cumulative frequency
351- 500	48	350.5- 500.5	48
501- 650	189	500.5- 650.5	237
651- 800	88	650.5- 800.5	325
801- 950	47	800.5- 950.5	372
951- 1100	28	950.5- 1100.5	400
Total	N = 400		

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$\frac{N}{4} = \frac{400}{4} = 100$,

Q_1 Class is 500.5 – 650.5

$l_1 = 500.5$, $m_1 = 48$, $f_1 = 189$, $c_1 = 150$

$$\therefore Q_1 = 500.5 + \frac{100 - 48}{189} \times 150$$

$$= 500.5 + \frac{52 \times 150}{189}$$

$$= 500.5 + 41.27$$

$$= 541.77$$

$$Q_3 = l_3 + \frac{3\frac{N}{4} - m_3}{f_3} \times c_3$$

$3\frac{N}{4} = 3 \times 100 = 300,$

Q3 Class is 650.5 – 800.5

$l_3 = 650.5, m_3 = 237, f_3 = 88, C_3 = 150$

$$\therefore Q_3 = 650.5 + \frac{300 - 237}{88} \times 150$$

$$= 650.5 + \frac{63 \times 150}{88}$$

$$= 650.5 + 107.39$$

$$= 757.89$$

$$\therefore Q.D = \frac{Q_3 - Q_1}{2}$$

$$= \frac{757.89 - 541.77}{2}$$

$$= \frac{216.12}{2}$$

$$= 108.06$$

$$\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{757.89 - 541.77}{757.89 + 541.77}$$

$$= \frac{216.12}{1299.66} = 0.1663$$

Merits and Demerits of Quartile Deviation Merits :

1. It is Simple to understand and easy to calculate
2. It is not affected by extreme values.
3. It can be calculated for data with open end classes also.

Demerits:

1. It is not based on all the items. It is based on two positional values Q_1 and Q_3 and ignores the extreme 50% of the items

2. It is not amenable to further mathematical treatment.
3. It is affected by sampling fluctuations.

Mean Deviation and Coefficient of Mean Deviation:

Mean Deviation:

The range and quartile deviation are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average. The mean deviation is measure of dispersion based on all items in a distribution.

Definition:

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored. According to Clark and Schekade,

"Average deviation is the average amount scatter of the items in a distribution from either the mean or the median, ignoring the signs of the deviations".

We usually compute mean deviation about any one of the three averages mean, median or mode. Some times mode may be ill defined and as such mean deviation is computed from mean and median. Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

Coefficient of mean deviation:

Mean deviation calculated by any measure of central tendency is an absolute measure. For the purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

Coefficient of mean deviation: $= \frac{\text{Mean deviation}}{\text{Mean or Median or Mode}}$

If the result is desired in percentage, the coefficient of mean deviation $= \frac{\text{Mean deviation}}{\text{Mean or Median or Mode}} \times 100$

Computation of mean deviation – Individual Series :

1. Calculate the average mean, median or mode of the series.
2. Take the deviations of items from average ignoring signs and denote these deviations by |D|.
3. Compute the total of these deviations, i.e., $\Sigma |D|$
4. Divide this total obtained by the number of items.

Symbolically: M.D. $= \frac{\Sigma|D|}{n}$

Example 6:

Calculate mean deviation from mean and median for the following data:

100,150,200,250,360,490,500,600,671 also calculate co-efficients of M.D.

Solution:

$$\text{Mean} = \bar{x} = \frac{\sum \mathbf{x}}{\mathbf{n}} = \frac{3321}{9} = 369$$

Now arrange the data in ascending order

100, 150, 200, 250, 360, 490, 500, 600, 671

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

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

$$= \text{Value of } 5^{th} \text{ item}$$

$$= 360$$

X	$\|D\| = \|x - \bar{x}\|$	$\|D\| = \|x - Md\|$
100	269	260
150	219	210
200	169	160
250	119	110
360	9	0
490	121	130
500	131	140
600	231	240
671	302	311
3321	1570	1561

M.D from mean $= \frac{\sum |D|}{n}$

$= \frac{1570}{9} = 174.44$

Co-efficient of M.D $= \frac{M.D}{\bar{x}}$

$= \frac{174.44}{369} = 0.47$

M.D from median $= \frac{\sum |D|}{n}$

$= \frac{1561}{9} = 173.44$

Co-efficient of M.D.$= \frac{M.D}{Median} = \frac{173.44}{360} = 0.48$

7.5.4 Mean Deviation – Discrete series:

Steps: 1. Find out an average (mean, median or mode)

2. Find out the deviation of the variable values from the average, ignoring signs and denote them by $|D|$

3. Multiply the deviation of each value by its respective frequency and find out the total $\sum f|D|$

4. Divide $\sum f |D|$ by the total frequencies N

Symbolically, M.D. = $\frac{\sum f|D|}{N}$

Example 7:

Compute Mean deviation from mean and median from the following data:

Height in cms	158	159	160	161	162	163	164	165	166
No. of persons	15	20	32	35	33	22	20	10	8

Also compute coefficient of mean deviation.

Solution:

Height X	No. of persons f	d= x- A A =162	fd	|D| = |X- mean|	f|D|
158	15	- 4	- 60	3.51	52.65
159	20	- 3	- 60	2.51	50.20
160	32	- 2	- 64	1.51	48.32
161	35	- 1	- 35	0.51	17.85
162	33	0	0	0.49	16.17
163	22	1	22	1.49	32.78
164	20	2	40	2.49	49.80
165	10	3	30	3.49	34.90
166	8	4	32	4.49	35.92
	195		- 95		338.59

$$\bar{x} = A + \frac{\sum fd}{N}$$

$$= 162 + \frac{-95}{195} = 162 - 0.49 = 161.51$$

$$\text{M.D.} = \frac{\sum f|D|}{N} = \frac{338.59}{195} = 1.74$$

Coefficient of M.D.= $\frac{M.D}{\bar{X}} = \frac{1.74}{161.51} = 0.0108$

Height x	No. of persons f	c.f.	$\lvert D \rvert = \lvert X - \text{Median} \rvert$	$f\lvert D \rvert$
158	15	15	3	45
159	20	35	2	40
160	32	67	1	32
161	35	102	0	0
162	33	135	1	33
163	22	157	2	44
164	20	177	3	60
165	10	187	4	40
166	8	195	5	40
	195			334

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

= Size of $\left(\frac{195+1}{2}\right)^{th}$ item

= Size of 98 th item

= 161

M.D = $\frac{\sum f\lvert D \rvert}{N} = \frac{334}{195} = 1.71$

Coefficient of M.D. = $\frac{M.D}{\text{Median}} = \frac{1.71}{161} = .0106$

7.5.5 Mean deviation-Continuous series:

The method of calculating mean deviation in a continuous series same as the discrete series.In continuous series we have to find out the mid points of the various classes and take deviation of these points from the average selected. Thus

$$M.D = \frac{\sum f\,|\,D\,|}{N}$$

Where D = m - average
M = Mid point

Example 8:

Find out the mean deviation from mean and median from the following series.

Age in years	No.of persons
0-10	20
10-20	25
20-30	32
30-40	40
40-50	42
50-60	35
60-70	10
70-80	8

Also compute co-efficient of mean deviation.

Solution:

X	m	f	$d = \frac{m - A}{c}$ (A=35,C=10)	fd	$\lvert D \rvert = \lvert m - \bar{x} \rvert$	$f\lvert D \rvert$
0-10	5	20	-3	-60	31.5	630.0
10-20	15	25	-2	-50	21.5	537.5
20-30	25	32	-1	-32	11.5	368.0
30-40	**35**	40	0	0	1.5	60.0
40-50	45	42	1	42	8.5	357.0
50-60	55	35	2	70	18.5	647.5
60-70	65	10	3	30	28.5	285.0
70-80	75	8	4	32	38.5	308.0
		212		32		3193.0

$$\bar{x} = A + \frac{\sum fd}{N} \times c$$

$$= 35 + \frac{32}{212} \times 10 \quad = 35 + \frac{320}{212} = 35 + 1.5 = 36.5$$

M.D. = $\frac{\sum f|D|}{N} = \frac{3193}{212} = 15.06$

Calculation of median and M.D. from median

X	m	f	c.f	|D| = |m-Md|	f |D|
0-10	5	20	20	32.25	645.00
10-20	15	25	45	22.25	556.25
20-30	25	32	77	12.25	392.00
30-40	35	40	117	2.25	90.00
40-50	45	42	159	7.75	325.50
50-60	55	35	194	17.75	621.25
60-70	65	10	204	27.75	277.50
70-80	75	8	212	37.75	302.00
				Total	3209.50

$\frac{N}{2} = \frac{212}{2} = 106$

l = 30, m = 77, f = 40, c = 10

Median = $l + \frac{\frac{N}{2} - m}{f} \times c$

$= 30 + \frac{106 - 77}{40} \times 10$

$= 30 + \frac{29}{4}$

$= 30 + 7.25 = 37.25$

M. D. = $\frac{\sum f|D|}{N}$

$= \frac{3209.5}{212} = 15.14$

Coefficient of M.D = $\frac{\text{M.D}}{\text{Median}}$

$= \frac{15.14}{37.25} = 0.41$

7.5.6 Merits and Demerits of M.D :

Merits:

1. It is simple to understand and easy to compute.
2. It is rigidly defined.
3. It is based on all items of the series.
4. It is not much affected by the fluctuations of sampling.
5. It is less affected by the extreme items.
6. It is flexible, because it can be calculated from any average.
7. It is better measure of comparison.

Demerits:

1. It is not a very accurate measure of dispersion.
2. It is not suitable for further mathematical calculation.
3. It is rarely used. It is not as popular as standard deviation.
4. Algebraic positive and negative signs are ignored. It is mathematically unsound and illogical.

Standard Deviation and Coefficient of variation:

Standard Deviation :

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

Definition:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean.

The standard deviation is denoted by the Greek letter σ (sigma)

Calculation of Standard deviation-Individual Series :

There are two methods of calculating Standard deviation in an individual series.

a) Deviations taken from Actual mean
b) Deviation taken from Assumed mean

a) Deviation taken from Actual mean:

This method is adopted when the mean is a whole number.

Steps:

1. Find out the actual mean of the series ($\bar{x}$)
2. Find out the deviation of each value from the mean $(x = X - \bar{X})$
3. Square the deviations and take the total of squared deviations $\sum x^2$
4. Divide the total ($\sum x^2$) by the number of observation $\left(\frac{\sum x^2}{n} \right)$

The square root of $\left(\frac{\sum x^2}{n} \right)$ is standard deviation.

Thus $\sigma = \sqrt{\left(\frac{\sum x^2}{n} \right)}$ or $\sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$

b) Deviations taken from assumed mean:

This method is adopted when the arithmetic mean is fractional value.

Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, We apply short –cut method; deviations are taken from an assumed mean. The formula is:

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N} \right)^2}$$

Where d-stands for the deviation from assumed mean = (X-A)

Steps:

1. Assume any one of the item in the series as an average (A)
2. Find out the deviations from the assumed mean; i.e., X-A denoted by d and also the total of the deviations $\sum d$
3. Square the deviations; i.e., d^2 and add up the squares of deviations, i.e, $\sum d^2$
4. Then substitute the values in the following formula:

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

Note: We can also use the simplified formula for standard deviation.

$$o = \frac{1}{n}\sqrt{n\sum d^2 - \left(\sum d\right)^2}$$

For the frequency distribution

$$o = \frac{c}{N}\sqrt{N\sum fd^2 - \left(\sum fd\right)^2}$$

Example 9:

Calculate the standard deviation from the following data.

14, 22, 9, 15, 20, 17, 12, 11

Solution:

Deviations from actual mean.

Values (X)	$X - \bar{X}$	$(X - \bar{X})^2$
14	-1	1
22	7	49
9	-6	36
15	0	0
20	5	25
17	2	4
12	-3	9
11	-4	16
120		140

$$\bar{X} = \frac{120}{8} = 15$$

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{140}{8}}$$

$$= \sqrt{17.5} = 4.18$$

Example 10:
The table below gives the marks obtained by 10 students in statistics. Calculate standard deviation.

Student Nos :	1	2	3	4	5	6	7	8	9	10
Marks :	43	48	65	57	31	60	37	48	78	59

Solution: (Deviations from assumed mean)

Nos.	Marks (x)	d=X-A (A=57)	d^2
1	43	-14	196
2	48	-9	81
3	65	8	64
4	57	0	0
5	31	-26	676
6	60	3	9
7	37	-20	400
8	48	-9	81
9	78	21	441
10	59	2	4
n = 10		$\sum d=-44$	$\sum d^2 =1952$

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

$$= \sqrt{\frac{1952}{10} - \left(\frac{-44}{10}\right)^2}$$

$$= \sqrt{195.2 - 19.36}$$

$$= \sqrt{175.84} = 13.26$$

Calculation of standard deviation:

Discrete Series:

There are three methods for calculating standard deviation in discrete series:

(a) Actual mean methods
(b) Assumed mean method
(c) Step-deviation method.

(a) Actual mean method:

Steps:

1. Calculate the mean of the series.
2. Find deviations for various items from the means i.e.,
   $x - \bar{x} = d$.
3. Square the deviations ($= d^2$) and multiply by the respective frequencies(f) we get fd^2
4. Total to product ($\sum fd^2$) Then apply the formula:

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f}}$$

If the actual mean in fractions, the calculation takes lot of time and labour; and as such this method is rarely used in practice.

(b) Assumed mean method:

Here deviation are taken not from an actual mean but from an assumed mean. Also this method is used, if the given variable values are not in equal intervals.

Steps:

1. Assume any one of the items in the series as an assumed mean and denoted by A.
2. Find out the deviations from assumed mean, i.e, X-A and denote it by d.
3. Multiply these deviations by the respective frequencies and get the $\sum fd$
4. Square the deviations (d^2).
5. Multiply the squared deviations ($d^{2)}$ by the respective frequencies (f) and get $\sum fd^{2\cdot}$
6. Substitute the values in the following formula:

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$$

Where $d = X - A$, $N = \sum f$.

Example 11:

Calculate Standard deviation from the following data.

X :	20	22	25	31	35	40	42	45
f :	5	12	15	20	25	14	10	6

Solution:

Deviations from assumed mean

x	f	d = x –A (A = 31)	d^2	fd	fd^2
20	5	-11	121	-55	605
22	12	-9	81	-108	972
25	15	-6	36	-90	540
31	20	0	0	0	0
35	25	4	16	100	400
40	14	9	81	126	1134
42	10	11	121	110	1210
45	6	14	196	84	1176
	N=107			$\sum fd=167$	$\sum fd^2$ =6037

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$$

$$= \sqrt{\frac{6037}{107} - \left(\frac{167}{107}\right)^2}$$

$$= \sqrt{56.42 - 2.44}$$

$$= \sqrt{53.98} = 7.35$$

(c) Step-deviation method:

If the variable values are in equal intervals, then we adopt this method.

Steps:

1. Assume the center value of the series as assumed mean A
2. Find out $d = \frac{x - A}{C}$, where C is the interval between each value
3. Multiply these deviations d' by the respective frequencies and get $\sum fd$
4. Square the deviations and get d^2
5. Multiply the squared deviation (d^2) by the respective frequencies (f) and obtain the total $\sum fd^2$

6. Substitute the values in the following formula to get the standard deviation.

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

Example 12:

Compute Standard deviation from the following data

Marks :	10	20	30	40	50	60
No.of students:	8	12	20	10	7	3

Solution:

Marks x	F	$d = \frac{x - 30}{10}$	fd	fd^2
10	8	-2	-16	32
20	12	-1	-12	12
30	20	0	0	0
40	10	1	10	10
50	7	2	14	28
60	3	3	9	27
	N=60		$\Sigma fd = 5$	$\Sigma fd^2 = 109$

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

$$= \sqrt{\frac{109}{60} - \left(\frac{5}{60}\right)^2} \times 10$$

$$= \sqrt{1.817 - 0.0069} \times 10$$

$$= \sqrt{1.8101} \times 10$$

$$= 1.345 \times 10$$

$$= 13.45$$

Calculation of Standard Deviation –Continuous series:

In the continuous series the method of calculating standard deviation is almost the same as in a discrete series. But in a continuous series, mid-values of the class intervals are to be found out. The step- deviation method is widely used.

The formula is,

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

$d = \frac{m - A}{C}$, C- Class interval.

Steps:

1. Find out the mid-value of each class.
2. Assume the center value as an assumed mean and denote it by A
3. Find out $d = \frac{m - A}{C}$
4. Multiply the deviations d by the respective frequencies and get Σfd
5. Square the deviations and get d^2
6. Multiply the squared deviations (d^2) by the respective frequencies and get $\sum fd^2$
7. Substituting the values in the following formula to get the standard deviation

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

Example 13:

The daily temperature recorded in a city in Russia in a year is given below.

Temperature C^0	No. of days
-40 to –30	10
-30 to –20	18
-20 to –10	30
-10 to 0	42
0 to 10	65
10 to 20	180
20 to 30	20
	365

Calculate Standard Deviation.

Solution:

Temperature	Mid value (m)	No. of days f	$d = \frac{m-(-5^n)}{10^n}$	fd	fd^2
-40 to -30	-35	10	-3	-30	90
-30 to -20	-25	18	-2	-36	72
-20 to -10	-15	30	-1	-30	30
-10 to - 0	-5	42	0	0	0
0 to 10	5	65	1	65	65
10 to 20	15	180	2	360	720
20 to 30	25	20	3	60	180
		N=365		Σfd = 389	Σfd^2 =1157

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times C$$

$$= \sqrt{\frac{1157}{365} - \left(\frac{389}{365}\right)^2} \times 10$$

$$= \sqrt{3.1699 - 1.1358} \times 10$$

$$= \sqrt{2.0341} \times 10$$

$$= 1.4262 \times 10$$

$$= 14.26° \text{ c}$$

Combined Standard Deviation:

If a series of N_1 items has mean $\overline{X}_1$ and standard deviation σ_1 and another series of N_2 items has mean $\overline{X}_2$ and standard deviation σ_2 , we can find out the combined mean and combined standard deviation by using the formula.

$$\overline{X}_{12} = \frac{N1\,\overline{X}_1 + N_2\,\overline{X}_2}{N_1 + N_2}$$

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

Where $d_1 = \overline{X}_1 - \overline{X}_{12}$

$$d_2 = \overline{X}_2 - \overline{X}_{12}$$

Example 14:

Particulars regarding income of two villages are given below.

	Village	
	A	**B**
No.of people	600	500
Average income	175	186
Standard deviation of income	10	9

Compute combined mean and combined Standard deviation.

Solution:

Given $N_1 = 600$, $\overline{X}_1 = 175$, $\sigma_1 = 10$

$N_2 = 500$, $\overline{X}_2 = 186$, $\sigma_2 = 9$

Combined mean $\overline{X}_{12} = \frac{N1\,\overline{X}_1 + N_2\,\overline{X}_2}{N_1 + N_2}$

$$= \frac{600 \times 175 + 500 \times 186}{600 + 500}$$

$$= \frac{105000 + 93000}{1100}$$

$$= \frac{198000}{1100} = 180$$

Combined Standard Deviation:

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$d_1 = \overline{X}_1 - \overline{X}_{12}$

$= 175 - 180$

$= -5$

$d_2 = \overline{X}_2 - \overline{X}_{12}$

$= 186 - 180$

$= 6$

$$\sigma_{12} = \sqrt{\frac{600 \times 100 + 500 \times 81 + 600 \times 25 + 500 \times 36}{600 + 500}}$$

$$= \sqrt{\frac{60000 + 40500 + 15000 + 18000}{1100}}$$

$$= \sqrt{\frac{133500}{1100}}$$

$= \sqrt{121.364}$

$= 11.02.$

Merits and Demerits of Standard Deviation:

Merits:

1. It is rigidly defined and its value is always definite and based on all the observations and the actual signs of deviations are used.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.
4. It is possible for further algebraic treatment.
5. It is less affected by the fluctuations of sampling and hence stable.
6. It is the basis for measuring the coefficient of correlation and sampling.

Demerits:

1. It is not easy to understand and it is difficult to calculate.
2. It gives more weight to extreme values because the values are squared up.
3. As it is an absolute measure of variability, it cannot be used for the purpose of comparison.

Coefficient of Variation :

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation.

The coefficient of variation is obtained by dividing the standard deviation by the mean and multiply it by 100. symbolically,

$$\text{Coefficient of variation (C.V)} = \frac{\sigma}{\bar{X}} \times 100$$

If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

Example 15:

In two factories A and B located in the same industrial area, the average weekly wages (in rupees) and the standard deviations are as follows:

Factory	Average	Standard Deviation	No. of workers
A	34.5	5	476
B	28.5	4.5	524

1. Which factory A or B pays out a larger amount as weekly wages?
2. Which factory A or B has greater variability in individual wages?

Solution:

Given $N_1 = 476$, $\bar{X}_1 = 34.5$, $\sigma_1 = 5$

$N_2 = 524$, $\bar{X}_2 = 28.5$, $\sigma_2 = 4.5$

1. Total wages paid by factory A

$= 34.5 \times 476$

$=$ Rs.16.422

Total wages paid by factory B

$= 28.5 \times 524$

$=$ Rs.14,934.

Therefore factory A pays out larger amount as weekly wages.

2. C.V. of distribution of weekly wages of factory A and B are

$$C.V.(A) = \frac{\sigma_1}{\bar{X}_1} \times 100$$

$$= \frac{5}{34.5} \times 100$$

$$= 14.49$$

$$C.V\ (B) = \frac{\sigma_2}{\bar{X}_2} \times 100$$

$$= \frac{4.5}{28.5} \times 100$$

$$= 15.79$$

Factory B has greater variability in individual wages, since C.V. of factory B is greater than C.V of factory A

Example 16:

Prices of a particular commodity in five years in two cities are given below:

Price in city A	Price in city B
20	10
22	20
19	18
23	12
16	15

Which city has more stable prices?

Solution:
Actual mean method

City A			City B		
Prices (X)	Deviations from X=20 dx	dx^2	Prices (Y)	Deviations from $\bar{Y}$ =15 dy	dy^2
20	0	0	10	-5	25
22	2	4	20	5	25
19	-1	1	18	3	9
23	3	9	12	-3	9
16	-4	16	15	0	0
$\Sigma x=100$	$\Sigma dx=0$	$\Sigma dx^2=30$	$\Sigma y=75$	$\Sigma dy=0$	$\Sigma dy^2=68$

City A: $\bar{X} = \frac{\Sigma x}{n} = \frac{100}{5} = 20$

$$\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma dx^2}{n}}$$

$$= \sqrt{\frac{30}{5}} = \sqrt{6} = 2.45$$

$$C.V(x) = \frac{\sigma_x}{\bar{x}} \times 100$$

$$= \frac{2.45}{20} \times 100$$

$$= 12.25\ \%$$

City B: $\bar{Y} = \frac{\Sigma y}{n} = \frac{75}{5} = 15$

$$\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} = \sqrt{\frac{\Sigma dy^2}{n}}$$

$$= \sqrt{\frac{68}{5}} = \sqrt{13.6} = 3.69$$

$$\text{C.V.}(y) = \frac{\sigma_y}{\bar{y}} \times 100$$

$$= \frac{3.69}{15} \times 100$$

$$= 24.6\ \%$$

City A had more stable prices than City B, because the coefficient of variation is less in City A.

Moments:

7.7.1 Definition of moments:

Moments can be defined as the arithmetic mean of various powers of deviations taken from the mean of a distribution. These moments are known as central moments.

The first four moments about arithmetic mean or central moments are defined below.

	Individual series	Discrete series
First moments about the mean; μ_1	$\frac{\Sigma(x - \bar{x})}{n} = 0$	$\frac{\sum f(x - \bar{x})}{N} = 0$
Second moments about the mean; μ_2	$\frac{\sum (x - \bar{x})^2}{n} = \sigma^2$	$\frac{\sum f(x - \bar{x})^2}{N}$
Third moments about the mean ; μ_3	$\frac{\sum (x - \bar{x})^3}{n}$	$\frac{\sum f(x - \bar{x})^3}{N}$
Fourth moment about the Mean ; μ_4	$\frac{\sum (x - \bar{x})^4}{n}$	$\frac{\sum f(x - \bar{x})^4}{N}$

μ is a Greek letter, pronounced as ‘ mu’ .

If the mean is a fractional value, then it becomes a difficult task to work out the moments. In such cases, we can calculate moments about a working origin and then change it into moments about the actual mean. The moments about an origin are known as raw moments.

The first four raw moments – individual series.

$$\mu_1 = \frac{\sum(X-A)}{N} = \frac{\sum d}{N} \qquad \mu_2 = \frac{\sum(X-A)^2}{N} = \frac{\sum d^2}{N}$$

$$\mu_3 = \frac{\sum(X-A)^3}{N} = \frac{\sum d^3}{N} \qquad \mu_4 = \frac{\sum(X-A)^4}{N} = \frac{\sum d^4}{N}$$

Where A – any origin, d=X-A

The first four raw moments – Discrete series (step – deviation method)

$$\mu'_1 = \frac{\sum fd'}{N} \times C \qquad \mu'_2 = \frac{\sum fd'^2}{N} \times C^2$$

$$\mu'_3 = \frac{\sum fd'^3}{N} \times C^3 \qquad \mu'_4 = \frac{\sum fd'^4}{N} \times C^4$$

Where $d = \frac{X-A}{C}$, A – origin , C – Common point

The first four raw Moments – Continuous series

$$\mu'_1 = \frac{\sum fd'}{N} \times C \qquad \mu'_2 = \frac{\sum fd'^2}{N} \times C^2$$

$$\mu'_3 = \frac{\sum fd'^3}{N} \times C^3 \qquad \mu'_4 = \frac{\sum fd'^4}{N} \times C^4$$

Where $d = \frac{m-A}{C}$ A – origin , C – Class internal

Relationship between Raw Moments and Central moments:

Relation between moments about arithmetic mean and moments about an origin are given below.

$$\mu_1 = \mu_1 - \mu_1 = 0$$

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

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

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

Example 17:

Calculate first four moments from the following data.

X :	0	1	2	3	4	5	6	7	8
F :	5	10	15	20	25	20	15	10	5

Solution:

X	f	fx	d=x- x (x-4)	fd	fd^2	fd^3	fd^4
0	5	0	-4	-20	80	-320	1280
1	10	10	-3	-30	90	-270	810
2	15	30	-2	-30	60	-120	240
3	20	60	-1	-20	20	-20	20
4	25	100	0	0	0	0	0
5	20	100	1	20	20	20	20
6	15	90	2	30	60	120	240
7	10	70	3	30	90	270	810
8	5	40	4	20	80	320	1280
	N =125	∑fx =500	∑d =0	∑fd =0	$\sum fd^2$ =500	$\sum fd^3$ =0	$\sum fd^4$ =4700

$$\bar{X} = \frac{\sum fx}{N} = \frac{500}{125} = 4$$

$$\mu_1 = \frac{\sum fd}{N} = \frac{0}{125} = 0 \qquad \mu_2 = \frac{\sum fd^2}{N} = \frac{500}{125} = 4$$

$$\mu_3 = \frac{\sum fd^3}{N} = \frac{0}{125} = 0 \qquad \mu_4 = \frac{\sum fd^4}{N} = \frac{4700}{125} = 37.6$$

Example 18:

From the data given below, first calculate the first four moments about an arbitrary origin and then calculate the first four moments about the mean.

X :	30-33	33-36	36-39	39-42	42-45	45-48
f :	2	4	26	47	15	6

Solution:

X	Midvalues (m)	f	$d = \frac{(m - 37.5)}{3}$	fd	fd^2	fd^3	fd 4
30-33	31.5	2	-2	-4	8	-16	32
33-36	34.5	4	-1	-4	4	-4	4
36-39	37.5	26	0	0	0	0	0
39-42	40.5	47	1	47	47	47	47
42-45	43.5	15	2	30	60	120	240
45-48	46.5	6	3	18	54	162	486
		N= 100		$\sum fd'$ =87	$\sum fd'^2$= 173	$\sum fd'^3$= 309	$\sum fd'^4$= 809

$$\mu_1 = \frac{\sum fd'}{N} \times c = \frac{87}{100} \times c = \frac{261}{100} = 2.61$$

$$\mu_2 = \frac{\sum fd'^2}{N} \times c^2 = \frac{173}{100} \times 9 = \frac{1557}{100} = 15.57$$

$$\mu_3 = \frac{\sum fd'^3}{N} \times c^3 = \frac{309}{100} \times 27 = \frac{8343}{100} = 83.43$$

$$\mu_4 = \frac{\sum fd'^4}{N} \times c^4 = \frac{809}{100} \times 81 = \frac{65529}{100} = 655.29$$

Moments about mean

$\mu_1 = 0$

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

$= 15.57 - (2.61)^2$

$= 15.57 - 6.81 = 8.76$

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

$= 83.43 - 3(2.61)(15.57) + 2(2.61)^3$

$= 83.43 - 121.9 + 35.56 = -2.91$

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

$= 665.29 - 4(83.43)(2.61) + 6(15.57)(2.61)^2 - 3(2.61)^4$

$= 665.29 - 871.01 + 636.39 - 139.214$

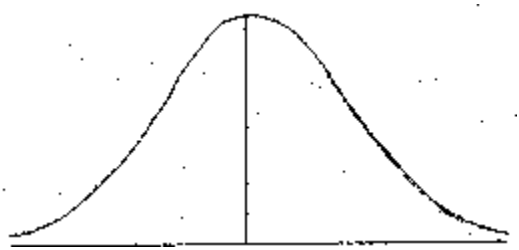
$= 291.454$

Skewness:

Meaning:

Skewness means ‘ lack of symmetry’ . We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data.If in a distribution mean = median = mode, then that distribution is known as symmetrical distribution. If in a distribution mean ≠ median ≠ mode , then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

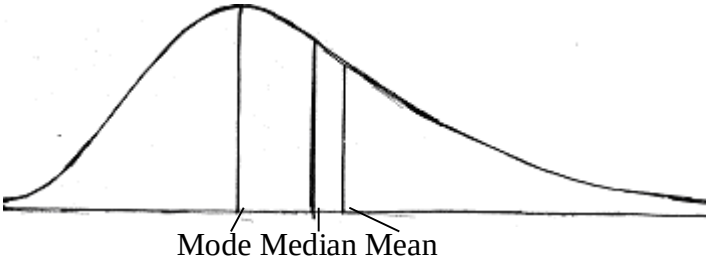
a) Symmetrical distribution:



Mean = Median = Mode

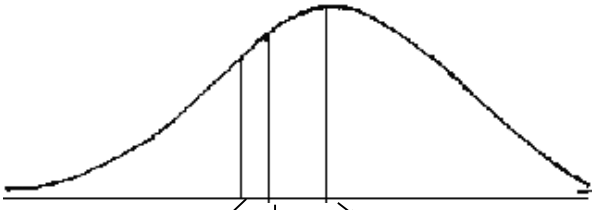
It is clear from the above diagram that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the center point of the curve.

b)Positively skewed distribution:



It is clear from the above diagram, in a positively skewed distribution, the value of the mean is maximum and that of the mode is least, the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the right hand side than they are on the left hand side.

c) Negatively skewed distribution:



It is clear from the above diagram, in a negatively skewed distribution, the value of the mode is maximum and that of the mean is least. The median lies in between the two. In the negatively skewed distribution the frequencies are spread out over a greater range of values on the left hand side than they are on the right hand side.

Measures of skewness:

The important measures of skewness are

(i) Karl – Pearason' s coefficient of skewness

(ii) Bowley' s coefficient of skewness

(iii)Measure of skewness based on moments

Karl – Pearson' s Coefficient of skewness:

According to Karl – Pearson, the absolute measure of skewness = mean – mode. This measure is not suitable for making valid comparison of the skewness in two or more distributions because the unit of measurement may be different in different series. To avoid this difficulty use relative measure of skewness called Karl – Pearson' s coefficient of skewness given by:

$$\text{Karl – Pearson' s Coefficient Skewness} = \frac{\text{Mean - Mode}}{S.D.}$$

In case of mode is ill – defined, the coefficient can be determined by the formula:

$$\text{Coefficient of skewness} = \frac{3(\text{Mean - Median})}{S.D.}$$

Example 18:

Calculate Karl – Pearson' s coefficient of skewness for the following data.

25, 15, 23, 40, 27, 25, 23, 25, 20

Solution:

Computation of Mean and Standard deviation :

Short – cut method.

Size	Deviation from A=25 D	d^2
25	0	0
15	-10	100
23	- 2	4
40	15	225
27	2	4
25	0	0
23	-2	4
25	0	0
20	- 5	25
N=9	$\sum d=-2$	$\sum d^2=362$

$$\text{Mean} = A + \frac{\sum d}{n}$$

$$= 25 + \frac{-2}{9}$$

$$= 25 - 0.22 = 24.78$$

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

$$= \sqrt{\frac{362}{9} - \left(\frac{-2}{9}\right)^2}$$

$$= \sqrt{40.22 - 0.05}$$

$$= \sqrt{40.17} = 6.3$$

Mode = 25, as this size of item repeats 3 times

Karl – Pearson' s coefficient of skewness

$$= \frac{\text{Mean - Mode}}{S.D.}$$

$$= \frac{24.78 - 25}{6.3}$$

$$= \frac{-0.22}{6.3}$$

$$= -0.03$$

Example 19:

Find the coefficient of skewness from the data given below

Size :	3	4	5	6	7	8	9	10
Frequency:	7	10	14	35	102	136	43	8

Solution:

Size	Frequency (f)	Deviation From A=6 (d)	d^2	fd	fd^2
3	7	-3	9	-21	63
4	10	-2	4	-20	40
5	14	-1	1	-14	14
6	35	0	0	0	0
7	102	1	1	102	102
8	136	2	4	272	544
9	43	3	9	129	387
10	8	4	16	32	128
	N=355			$\sum fd=480$	$\sum fd^2=1278$

$$\text{Mean} = A + \frac{\sum fd}{N} \qquad \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

$$= 6 + \frac{480}{355} \qquad = \sqrt{\frac{1278}{355} - \left(\frac{480}{355}\right)^2}$$

$$= 6 + 1.35 \qquad = \sqrt{3.6 - 1.82}$$

$$= 7.35 \qquad = \sqrt{1.78} = 1.33$$

Mode = 8

$$\text{Coefficient of skewness} = \frac{\text{Mean - Mode}}{S.D.}$$

$$= \frac{7.35 - 8}{1.33} = \frac{0.65}{1.33} = -0.5$$

Example 20:

Find Karl – Pearson' s coefficient of skewness for the given distribution:

X :	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
F :	2	5	7	13	21	16	8	3

Solution:

Mode lies in 20-25 group which contains the maximum frequency

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times C$$

l =20, f_1=21, f_0=13, f_2=16, C=5

$$\text{Mode} = 20 + \frac{21-13}{2 \times 21 - 13 - 16} \times 5$$

$$= 20 + \frac{8 \times 5}{42 - 29}$$

$$= 20 + \frac{40}{13} = 20 + 3.08 = 23.08$$

Computation of Mean and Standard deviation

X	Mid – point M	Frequency f	Deviations $d = \frac{m - 22.5}{5}$	fd	d^2	fd'^2
0-5	2.5	2	-4	-8	16	32
5-10	7.5	5	-3	-15	9	45
10-15	12.5	7	-2	-14	4	28
15-20	17.5	13	-1	-13	1	13
20-25	22.5	21	0	0	0	0
25-30	27.5	16	1	16	1	16
30-35	32.5	8	2	16	4	32
35-40	37.5	3	3	9	9	27
		N=75		$\sum fd$ = -9		$\sum fd^2$= 193

$$\text{Mean} = A + \frac{\sum fd}{N} \times c$$

$$= 22.5 + \frac{-9}{75} \times 5$$

$$= 22.5 - \frac{45}{75}$$

$$= 22.5 - 0.6 = 21.9$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times c$$

$$= \sqrt{\frac{193}{75} - \left(\frac{-9}{75}\right)^2} \times 5$$

$$= \sqrt{2.57 - 0.0144} \times 5$$

$$= \sqrt{2.5556} \times 5$$

$$= 1.5986 \times 5 = 7.99$$

Karl – Pearson' s coefficient of skewness

$$= \frac{\text{Mean - Mode}}{S.D.}$$

$$= \frac{21.9 - 23.08}{7.99}$$

$$= \frac{-1.18}{7.99} = -0.1477$$

7.10.2 Bowley' s Coefficient of skewness:

In Karl – Pearson' s method of measuring skewness the whole of the series is needed. Prof. Bowley has suggested a formula based on relative position of quartiles. In a symmetrical distribution, the quartiles are equidistant from the value of the median; ie.,

Median – Q_1 = Q_3 – Median. But in a skewed distribution, the quartiles will not be equidistant from the median. Hence Bowley has suggested the following formula:

$$\text{Bowley' s Coefficient of skewness (sk)} = \frac{Q_3 + Q_1 - 2\ \text{Median}}{Q_3 - Q_1}$$

Example 21:

Find the Bowley' s coefficient of skewness for the following series.

2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22

Solution:

The given data in order

2, 4, 6, 10, 12, 14, 16, 18, 20, 22

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

= size of $\left(\frac{11+1}{4}\right)^{th}$ item

= size of 3rd item = 6

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

= size of $3\left(\frac{11+1}{4}\right)^{th}$ item

= size of 9^{th} item

= 18

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

= size of $\left(\frac{11+1}{2}\right)$ th item

= size of 6^{th} item

= 12

Bowley' s coefficient skewness = $\frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$

$= \frac{18 + 6 - 2\times 12}{18 - 6} = 0$

Since sk = 0, the given series is a symmetrical data.

Example 22:

Find Bowley' s coefficient of skewness of the following series.

Size :	4	4.5	5	5.5	6	6.5	7	7.5	8
f :	10	18	22	25	40	15	10	8	7

Solution:

Size	f	c.f
4	10	10
4.5	18	28
5	22	50
5.5	25	75
6	40	115
6.5	15	130
7	10	140
7.5	8	148
8	7	155

Q_1 = Size of $\left(\frac{N+1}{4}\right)^{th}$ item

= Size of $\left(\frac{155+1}{4}\right)^{th}$ item

= Size of 39^{th} item

= 5

Q_2 = Median = Size of $\left(\frac{N+1}{2}\right)^{th}$ item

= Size of $\left(\frac{155+1}{2}\right)$ th item

= Size of 78^{th} item

= 6

Q_3 = Size of $3\left(\frac{N+1}{4}\right)^{th}$ item

= Size of $3\left(\frac{155+1}{4}\right)^{th}$ item

= Size of 117^{th} item = 6.5

Bowley' s Coefficient Skewness $= \dfrac{Q_3 + Q_1 - 2\ \text{Median}}{Q_3 - Q_1}$

$$= \frac{6.5 + 5 - 2 \times 6}{6.5 - 5}$$
$$= \frac{11.5 - 12}{1.5} = \frac{0.5}{1.5}$$
$$= -0.33$$

Example 23:

Calculate the value of the Bowley' s coefficient of skewness from the following series.

Wages : (Rs)	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No.of Persons :	1	3	11	21	43	32	9

Solution:

Wages(Rs)	F	c.f
10-20	1	1
20-30	3	4
30-40	11	15
40-50	21	36
50-60	43	79
60-70	32	111
70-80	9	120
	N=120	

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$$\frac{N}{4} = \frac{120}{4} = 30$$

Q_1class = 40-50

l_1= 40, m_1=15, f_1=21, c_1=10

$$\therefore Q_1 = 40 + \frac{30 - 15}{21} \times 10$$

$$= 40 + \frac{150}{21}$$

$$= 40 + 7.14$$

$$= 47.14$$

$$Q_2 = \text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$\frac{N}{2} = \frac{120}{2} = 60$$

Medianal class = 50 – 60

l= 50 , m=36, f = 43, c=10

$$\text{median} = 50 + \frac{60 - 36}{43} \times 10$$

$$= 50 + \frac{240}{43}$$

$$= 50 + 5.58$$

$$= 55.58$$

$$Q_3 = l_3 + \frac{3\frac{N}{4} - m}{f_3} \times c_3$$

$$3\frac{N}{4} = 3 \times \frac{120}{4} = 90$$

Q_3 class = 60 –70

l_3=60, m_3=79, f_3=32, c_3=10

$$\therefore Q_3 = 60 + \frac{90 - 79}{32} \times 10$$

$$= 60 + \frac{110}{32}$$

$$= 60 + 3.44$$

$$= 63.44$$

Bowley' s Coefficient of skewness $= \dfrac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$

$= \dfrac{63.44 + 47.14 - 2 \times 55.58}{63.44 - 47.14}$

$= \dfrac{110.58 - 111.16}{16.30}$

$= \dfrac{-0.58}{16.30}$

$= -0.0356$

7.10.3 Measure of skewness based on moments:

The measure of skewness based on moments is denoted by β_1 and is given by:

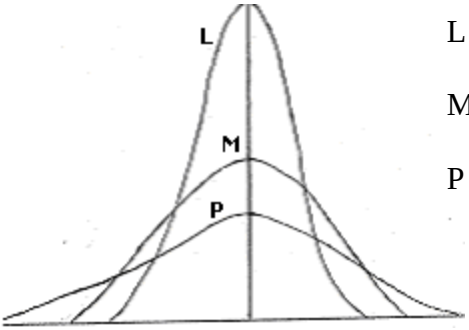
$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$ If μ_3 is negative, then β_1 is negative

Kurtosis:

The expression ' Kurtosis' is used to describe the peakedness of a curve.

The three measures – central tendency, dispersion and skewness describe the characteristics of frequency distributions. But these studies will not give us a clear picture of the characteristics of a distribution.

As far as the measurement of shape is concerned, we have two characteristics – skewness which refers to asymmetry of a series and kurtosis which measures the peakedness of a normal curve. All the frequency curves expose different degrees of flatness or peakedness. This characteristic of frequency curve is termed as kurtosis. Measure of kurtosis denote the shape of top of a frequency curve. Measure of kurtosis tell us the extent to which a distribution is more peaked or more flat topped than the normal curve, which is symmetrical and bell-shaped, is designated as Mesokurtic. If a curve is relatively more narrow and peaked at the top, it is designated as Leptokurtic. If the frequency curve is more flat than normal curve, it is designated as platykurtic.



L = Lepto Kurtic

M = Meso Kurtic

P = Platy Kurtic

Measure of Kurtosis:

The measure of kurtosis of a frequency distribution based moments is denoted by β_2 and is given by

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

If $\beta_2 = 3$, the distribution is said to be normal and the curve is mesokurtic.

If $\beta_2 > 3$, the distribution is said to be more peaked and the curve is leptokurtic.

If $\beta_2 < 3$, the distribution is said to be flat topped and the curve is platykurtic.

Example 24:

Calculate β_1 and β_2 for the following data.

X :	0	1	2	3	4	5	6	7	8
F :	5	10	15	20	25	20	15	10	5

Solution:

[**Hint:** Refer Example of page 172 and get the values of first four central moments and then proceed to find β_1 and β_2]

$\mu_1 = 0$ $\qquad$ $\mu_2 = \frac{\sum fd^2}{N} = \frac{500}{125} = 4$

$\mu_3 = \frac{\sum fd^3}{N} = 0$ $\qquad$ $\mu_4 = \frac{\sum fd^4}{N} = \frac{4700}{125} = 37.6$

$$\therefore \beta_1 = \frac{\mu^2_3}{\mu^3_2} = \frac{0}{64} = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{37.6}{4^2}$$

$$= \frac{37.6}{16} = 2.35$$

The value of β_2 is less than 3, hence the curve is platykurtic.

Example 25:

From the data given below, calculate the first four moments about an arbitrary origin and then calculate the first four central moments.

X :	30-33	33-36	36-39	39-42	42-45	45-48
f :	2	4	26	47	15	6

Solution:

[**Hint:** Refer Example 18 of page 172 and get the values of first four moments about the origin and the first four moments about the mean. Then using these values find the values of β_1 and β_2.]

$\mu_1 = 0$, $\mu_2 = 8.76$ $\mu_3 = -2.91$, $\mu_4 = 291.454$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad \beta_1 = \frac{(-2.91)^2}{(8.76)^3} = \frac{8.47}{672.24} = 0.0126$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \qquad \beta_2 = \frac{291.454}{(8.76)^2} = 3.70$$

Since $\beta_2 > 3$, the curve is leptokurtic.

Exercise – 7

I. Choose the best answer:

1. Which of the following is a unitless measure of dispersion?
   (a) Standard deviation (b) Mean deviation
   (c) Coefficient of variation (d) Range
2. Absolute sum of deviations is minimum from
   (a) Mode (b) Median
   (c) Mean (d) None of the above

3.In a distribution S.D = 6. All observation multiplied by 2 would give the result to S.D is

(a) 12 (b)6 (c) 18 (d) $\sqrt{6}$

4. The mean of squared deviations about the mean is called

(a) S.D (b) Variance (c) M.D (d) None

5. 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) 24

6. Quartile deviation is equal to

(a) Inter quartile range (b) double the inter quartile range

(c) Half of the inter quartile range (d) None of the above

7. Which of the following measures is most affected by extreme values

(a) S.D (b) Q.D (c) M.D (d) Range

8. Which measure of dispersion ensures highest degree of reliability?

(a) Range (b) Mean deviation (c) Q.D (d) S.D

9. For a negatively skewed distribution, the correct inequality is

(a) Mode < median (b) mean < median

(c) mean < mode (d) None of the above

10.In case of positive skewed distribution, the extreme values lie in the

(a) Left tail (b) right tail (c) Middle (d) any where

II. Fill in the blanks:

11. Relative measure of dispersion is free from _____

12. _____is suitable for open end distributions.

13. The mean of absolute deviations from an average is called _____

14. Variance is 36, the standard deviation is _____

15. The standard deviation of the five observations 5, 5,5,5,5 is _____

16. The standard deviation of 10 observation is 15. If 5 is added to each observations the vale of new standard deviation is _____

17. The second central moment is always a _____

18. If $\bar{x} = 50$, mode = 48, $\sigma = 20$, the coefficient of skewness shall be _____
19. In a symmetrical distribution the coefficient of skewness is _____
20. If $\beta_2 = 3$ the distribution is called _____

III. Answer the following

21. What do you understand by dispersion? What purpose does a measure of dispersion serve?
22. Discuss various measures of dispersion
23. Mention the characteristics of a good measure of dispersion.
24. Define Mean deviation and coefficient of mean deviation.
25. Distinguish between Absolute and relative measures of dispersion
26. List out merits and demerits of Mean deviation
27. Define quartile deviation and coefficient of quartile deviation.
28. Mention all the merits and demerits of quartile deviation
29. Define standard deviation. Also mention its merits and demerits
30. What is coefficient of variation? What purpose does it serve?
31. What do you understand by skewness. What are the various measures of skewness
32. What do you understand by kurtosis? What is the measure of measuring kurtosis?
33. Distinguish between skewness and kurtosis and bring out their importance in describing frequency distribution.
34. Define moments. Also distinguish between raw moments and central moments.
35. Mention the relationship between raw moments and central moments for the first four moments.
36. Compute quartile deviation from the following data.

Height in inches:	58	59	60	61	62	63	64	65	66
No.of students :	15	20	32	35	33	22	20	10	8

37. Compute quartile deviation from the following data :

Size :	4-8	8-12	12-16	16-20	20-24	24-28	28-32	32-36	36-40
Frequency:	6	10	18	30	15	12	10	6	2

38. Calculate mean deviation from mean from the following data:

X :	2	4	6	8	10
f :	1	4	6	4	1

39.Calculate mean deviation from median

Age	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55
No. of People	9	16	12	26	14	12	6	5

40.Calculate the S.D of the following

Size	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

41. Calculate S.D from the following series

Class interval	5-15	15-25	25-35	35-45	45-55
Frequency	8	12	15	9	6

42.Find out which of the following batsmen is more consistent in scoring.

Batsman A	5	7	16	27	39	53	56	61	80	101	105
Batsman B	0	4	16	21	41	43	57	78	83	93	95

43.Particulars regarding the income of two villages are given below:

	Village A	Village B
Number of people	600	500
Average income (in Rs)	175	186
Variance of income (in Rs)	100	81

In which village is the variation in income greater)

44. From the following table calculate the Karl – Pearson's coefficient of skewness

Daily Wages(in Rs):	150	200	250	300	350	400	450
No. of People	3	25	19	16	4	5	6

45. Compute Bowley' s coefficient of skewness from the following data:

Size	5-7	8-10	11-13	14-16	17-19
Frequency	14	24	38	20	4

46. Using moments calculate β1 and β2 from the following data:

Daily wages	70-90	90-110	110-130	130-150	150-170
No. of workers	8	11	18	9	4

IV. Suggested Activity

Select any two groups of any size from your class calculate mean, S.D and C.V for statistics marks. Find which group is more consistent.

Answers

I. 1. (c) 2. (c) 3. (a) 4. (b) 5. (b)
6. (c) 7. (d) 8. (d) 9. (c) 10.(b)

II.

11. units 12. Q.D 13. M.D 14. 6 15. Zero
16. 15 17. Variance 18. 0.1 19. zero
20. Mesokurtic

III.

36. Q.D = 1.5 37. Q.D = 5.2085 38. M.D = 1.5
39. M.D = 7.35 40. S.D = 1.67 41. S.D = 12.3

42. S.D of A = 67.06
S.D of B = 68.8

43. C.V.A = 5.71% ; C.V.B = 4.84 %
44. Sk = 0.88 45. Sk = – 0.13
46. $\beta_1 = 0.006$ $\beta_2 = 2.305$

8. CORRELATION

Introduction:

The term correlation is used by a common man without knowing that he is making use of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

The study related to the characteristics of only variable such as height, weight, ages, marks, wages, etc., is known as univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bi-Variate Analysis. Some times the variables may be inter-related. In health sciences we study the relationship between blood pressure and age, consumption level of some nutrient and weight gain, total income and medical expenditure, etc., The nature and strength of relationship may be examined by correlation and Regression analysis.

Thus Correlation refers to the relationship of two variables or more. (e-g) relation between height of father and son, yield and rainfall, wage and price index, share and debentures etc.

Correlation is statistical Analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. Price and supply, income and expenditure are correlated.

Definitions:

1. Correlation Analysis attempts to determine the degree of relationship between variables- Ya-Kun-Chou.
2. Correlation is an analysis of the covariation between two or more variables.- A.M.Tuttle.

Correlation expresses the inter-dependence of two sets of variables upon each other. One variable may be called as (subject)

independent and the other relative variable (dependent). Relative variable is measured in terms of subject.

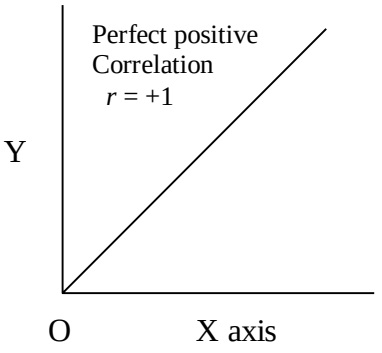
Uses of correlation:

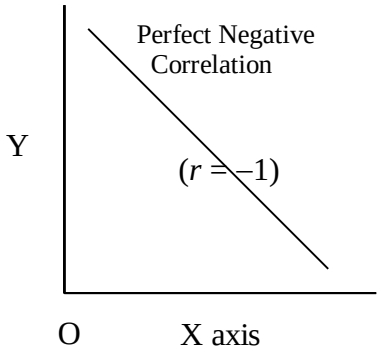
1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity etc. Businessmen estimates costs, sales, price etc. using correlation.
3. It is helpful in measuring the degree of relationship between the variables like income and expenditure, price and supply, supply and demand etc.
4. Sampling error can be calculated.
5. It is the basis for the concept of regression.

Scatter Diagram:

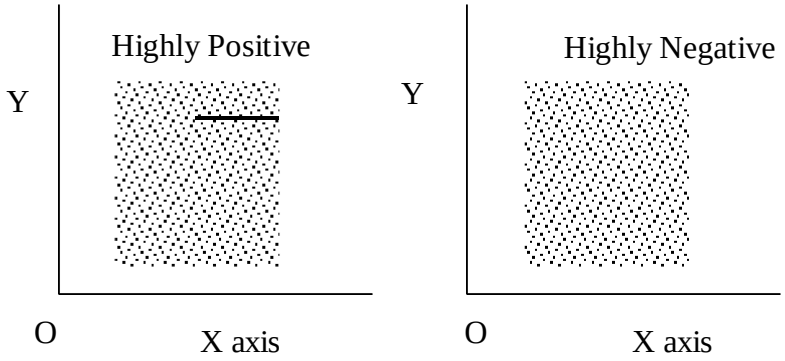
It is the simplest method of studying the relationship between two variables diagrammatically. One variable is represented along the horizontal axis and the second variable along the vertical axis. For each pair of observations of two variables, we put a dot in the plane. There are as many dots in the plane as the number of paired observations of two variables. The direction of dots shows the scatter or concentration of various points. This will show the type of correlation.

1. If all the plotted points form a straight line from lower left hand corner to the upper right hand corner then there is Perfect positive correlation. We denote this as $r = +1$

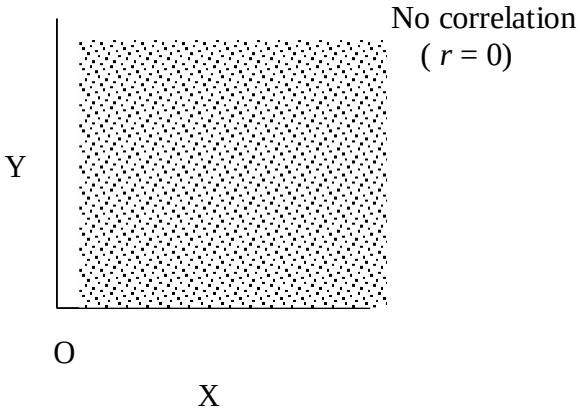




1. If all the plotted dots lie on a straight line falling from upper left hand corner to lower right hand corner, there is a perfect negative correlation between the two variables. In this case the coefficient of correlation takes the value $r = -1$.
2. If the plotted points in the plane form a band and they show a rising trend from the lower left hand corner to the upper right hand corner the two variables are highly positively correlated.



1. If the points fall in a narrow band from the upper left hand corner to the lower right hand corner, there will be a high degree of negative correlation.
2. If the plotted points in the plane are spread all over the diagram there is no correlation between the two variables.



Merits:

1. It is a simplest and attractive method of finding the nature of correlation between the two variables.
2. It is a non-mathematical method of studying correlation. It is easy to understand.
3. It is not affected by extreme items.
4. It is the first step in finding out the relation between the two variables.
5. We can have a rough idea at a glance whether it is a positive correlation or negative correlation.

Demerits:

By this method we cannot get the exact degree or correlation between the two variables.

Types of Correlation:

Correlation is classified into various types. The most important ones are

i) Positive and negative.
ii) Linear and non-linear.
iii) Partial and total.
iv) Simple and Multiple.

Positive and Negative Correlation:

It depends upon the direction of change of the variables. If the two variables tend to move together in the same direction (ie) an increase in the value of one variable is accompanied by an increase in the value of the other, (or) a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive or direct correlation. Price and supply, height and weight, yield and rainfall, are some examples of positive correlation.

If the two variables tend to move together in opposite directions so that increase (or) decrease in the value of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative (or) inverse correlation. Price and demand, yield of crop and price, are examples of negative correlation.

Linear and Non-linear correlation:

If the ratio of change between the two variables is a constant then there will be linear correlation between them.

Consider the following.

X	2	4	6	8	10	12
Y	3	6	9	12	15	18

Here the ratio of change between the two variables is the same. If we plot these points on a graph we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curvi-linear (or) non-linear correlation. The graph will be a curve.

Simple and Multiple correlation:

When we study only two variables, the relationship is simple correlation. For example, quantity of money and price level, demand and price. But in a multiple correlation we study more than two variables simultaneously. The relationship of price, demand and supply of a commodity are an example for multiple correlation.

Partial and total correlation:

The study of two variables excluding some other variable is called **Partial correlation**. For example, we study price and demand eliminating supply side. In total correlation all facts are taken into account.

Computation of correlation:

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure of correlation (or) correlation coefficient and it is denoted by ' r' .

Co-variation:

The covariation between the variables x and y is defined as

Cov(x,y) = $\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$ where $\bar{x}, \bar{y}$ are respectively means of x and y and ' n' is the number of pairs of observations.

Karl pearson' s coefficient of correlation:

Karl pearson, a great biometrician and statistician, suggested a mathematical method for measuring the magnitude of linear relationship between the two variables. It is most widely used method in practice and it is known as pearsonian coefficient of correlation. It is denoted by ' r' . The formula for calculating ' r' is

(i) $r = \dfrac{Cov(x, y)}{\sigma_x . \sigma_y}$ where σ_x , σ_y are S.D of x and y respectively.

(ii) $r = \dfrac{\sum xy}{n\,\sigma_x\,\sigma_y}$

(iii) $r = \dfrac{\Sigma XY}{\sqrt{\sum X^2 . \sum Y^2}}$, $X = x - \bar{x}$, $Y = y - \bar{y}$

when the deviations are taken from the actual mean we can apply any one of these methods. Simple formula is the third one.

The third formula is easy to calculate, and it is not necessary to calculate the standard deviations of x and y series respectively.

Steps:

1. Find the mean of the two series x and y.
2. Take deviations of the two series from x and y. $X = x - \bar{x}$, $Y = y - \bar{y}$
3. Square the deviations and get the total, of the respective squares of deviations of x and y and denote by ΣX^2 , ΣY^2 respectively.
4. Multiply the deviations of x and y and get the total and Divide by n. This is covariance.
5. Substitute the values in the formula.

$$r = \frac{cov(x, y)}{\sigma x . \sigma y} = \frac{\Sigma(x - \bar{x})\,(y - \bar{y}) / n}{\sqrt{\dfrac{\Sigma(x - \bar{x})^2}{n}} . \sqrt{\dfrac{\Sigma(y - \bar{y})^2}{n}}}$$

The above formula is simplified as follows

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 . \Sigma Y^2}}, \quad X = x - \bar{x}, \; Y = y - \bar{y}$$

Example 1:

Find Karl Pearson' s coefficient of correlation from the following data between height of father (x) and son (y).

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Comment on the result.

Solution:

x	Y	$X = x - \bar{x}$ X = x – 67	X^2	$Y = y - \bar{y}$ Y = y - 68	Y^2	XY
64	66	-3	9	-2	4	6
65	67	-2	4	-1	1	2
66	65	-1	1	-3	9	3
67	68	0	0	0	0	0
68	70	1	1	2	4	2
69	68	2	4	0	0	0
70	72	3	9	4	16	12
469	476	0	28	0	34	25

$\bar{x} = \frac{469}{7} = 67$; $y = \frac{476}{7} = 68$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 . \Sigma Y^2}} = \frac{25}{\sqrt{28 \times 34}} = \frac{25}{\sqrt{952}} = \frac{25}{30.85} = 0.81$$

Since r = + 0.81, the variables are highly positively correlated. (ie) Tall fathers have tall sons.

Working rule (i)

We can also find *r* with the following formula

We have $r = \frac{Cov(x, y)}{\sigma_x . \sigma_y}$

$$Cov(x,y) = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n} = \frac{\Sigma(xy + \bar{x}\bar{y} - \bar{y}x - \bar{x}\bar{y})}{n}$$

$$= \frac{\Sigma xy}{n} - \frac{\bar{y}\Sigma x}{n} - \frac{\bar{x}\Sigma y}{n} + \frac{\Sigma \bar{x}\bar{y}}{n}$$

$$\text{Cov}(x,y) = \frac{\Sigma xy}{n} - \cancel{\bar{y}\bar{x}} - \bar{x}\bar{y} + \cancel{\bar{x}\bar{y}} = \frac{\Sigma xy}{n} - \bar{x}\bar{y}$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - \bar{x}^2, \quad \sigma_y^2 = \frac{\Sigma y^2}{n} - \bar{y}^2$$

Now $r = \frac{Cov(x,y)}{\sigma_x . \sigma_y}$

$$r = \frac{\frac{\Sigma xy}{n} - \bar{x}\bar{y}}{\sqrt{\left(\frac{\Sigma x^2}{n} - \bar{x}^2\right)} \cdot \sqrt{\left(\frac{\Sigma y^2}{n} - \bar{y}^2\right)}}$$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Note: In the above method we need not find mean or standard deviation of variables separately.

Example 2:

Calculate coefficient of correlation from the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

x	y	x^2	y^2	xy
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
45	108	285	1356	597

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$r = \frac{9 \times 597 - 45 \times 108}{\sqrt{\left(9 \times 285 - (45)^2\right).(9 \times 1356 - (108)^2)}}$$

$$r = \frac{5373 - 4860}{\sqrt{(2565 - 2025)(.12204 - 11664)}}$$

$$= \frac{513}{\sqrt{540 \times 540}} = \frac{513}{540} = 0.95$$

Working rule (ii) (shortcut method)

We have $r = \frac{Cov(x, y)}{\sigma_x . \sigma_y}$

where $Cov(x,y) = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n}$

Take the deviation from x as x – A and the deviation from y as y – B

$$Cov(x,y) = \frac{\Sigma\ [(x - A) - (\bar{x} - A)]\ [(y - B) - (\bar{y} - B)]}{n}$$

$$= \frac{1}{n}\Sigma\ [(x - A)(y - B) - (x - A)(\bar{y} - B) - (\bar{x} - A)(y - B) + (\bar{x} - A)(\bar{y} - B)]$$

$$= \frac{1}{n}\Sigma\ [(x - A)(y - B) - (\bar{y} - B)\frac{\Sigma(x - A)}{n} - (\bar{x} - A)\frac{\Sigma(y - B)}{n} + \frac{\Sigma(\bar{x} - A)(\bar{y} - B)}{n}$$

$$= \frac{\Sigma(x - A)(y - B)}{n} - (y - B)(\bar{x} - \frac{nA}{n}) - (\bar{x} - A)(\bar{y} - \frac{nB}{n}) + (\bar{x} - A)(\bar{y} - B)$$

$$= \frac{\Sigma(x - A)(y - B)}{n} - (\bar{y} - B)(\bar{x} - A)$$

$$- \cancel{(\bar{x} - A)(\bar{y} - B)} + \cancel{(\bar{x} - A)(\bar{y} - B)}$$

$$= \frac{\Sigma(x - A)(y - B)}{n} - (\bar{x} - A)(\bar{y} - B)$$

Let $x - A = u$; $y - B = v$; $\bar{x} - A = \bar{u}$; $\bar{y} - B = \bar{v}$

$$\therefore \text{Cov}(x,y) = \frac{\Sigma uv}{n} - \bar{u}\bar{v}$$

$$\sigma_x^2 = \frac{\Sigma u^2}{n} - \bar{u}^2 = \sigma u^2$$

$$\sigma_y^2 = \frac{\Sigma v^2}{n} - \bar{v}^2 = \sigma v^2$$

$$\therefore r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{\left[n\Sigma u^2 - (\Sigma u)^2\right].\left[(n\Sigma v^2) - (\Sigma v)^2\right.}}$$

Example 3:

Calculate Pearson' s Coefficient of correlation.

X	45	55	56	58	60	65	68	70	75	80	85
Y	56	50	48	60	62	64	65	70	74	82	90

X	Y	u = x-A	v = y-B	u^2	v^2	uv
45	56	-20	-14	400	196	280
55	50	-10	-20	100	400	200
56	48	-9	-22	81	484	198
58	60	-7	-10	49	100	70
60	62	-5	-8	25	64	40
65	64	0	-6	0	36	0
68	65	3	-5	9	25	-15
70	70	5	0	25	0	0
75	74	10	4	100	16	40
80	82	15	12	225	144	180
85	90	20	20	400	400	400
		2	-49	1414	1865	1393

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{[n\Sigma u^2 - (\Sigma u^2)]\,[n\Sigma v^2 - (\Sigma v)^2]}}$$

$$r = \frac{11 \times 1393 - 2 \times (-49)}{\sqrt{(1414 \times 11 - (2)^2) \times (1865 \times 11 - (-49)^2)}}$$

$$= \frac{15421}{\sqrt{15550 \times 18114}} = \frac{15421}{16783.11} = +0.92$$

Correlation of grouped bi-variate data:

When the number of observations is very large, the data is classified into two way frequency distribution or correlation table. The class intervals for ' y' are in the column headings and for ' x' in the stubs. The order can also be reversed. The frequencies for each cell of the table are obtained. The formula for calculation of correlation coefficient ' r' is

$$r = \frac{\text{cov}(x, y)}{\sigma x, \sigma y} \quad \text{Where cov(x,y)} = \frac{\Sigma f (x - \bar{x})(y - \bar{y})}{N}$$

$$= \frac{\Sigma fxy}{N} - \bar{x}\,\bar{y}$$

$$\sigma x^2 = \frac{\Sigma fx^2}{N} - \bar{x}^2 \;;\; \sigma y^2 = \frac{\Sigma fy^2}{N} - \bar{y}^2$$

N – total frequency

$$r = \frac{N\Sigma fxy - (\Sigma fx)(\Sigma fy)}{\sqrt{[N\Sigma fx^2 - (\Sigma fx)^2].[N\Sigma fy^2 - (\Sigma fy)^2]}}$$

Theorem: The correlation coefficient is not affected by change of origin and scale.

If $u = \frac{x - A}{c}$; $v = \frac{y - B}{d}$ then $r_{xy} = r_{uv}$

Proof:

$$u = \frac{x - A}{}$$

c

$cu = x - A$

$x = cu + A$

$\bar{x} = c\,\bar{u} + A$

$$v = \frac{y - B}{d}$$

$vd = y - B$

$y = B + vd \quad \bar{y} = [B + \bar{v}d]$

$\sigma_x = c\sigma_u ; \quad \sigma_y = d\sigma_v$

$$r_{xy} = \frac{\text{cov}(x, y)}{\sigma_x, \sigma_y}$$

$$\text{cov}(x,y) = \frac{\Sigma f (x - \bar{x})(y - \bar{y})}{n}$$

$$\frac{1}{n}\Sigma f[(cu+A) - (c\bar{u}+A)][(dv+B) - (d\bar{v}+B)]$$

$$= \frac{1}{n}\Sigma f\left[cu - c\bar{u}\right]\left[(dv - d\bar{v})\right]$$

$$= \frac{1}{N}\Sigma f\left[c(u - \bar{u})\right]\left[d(v - \bar{v})\right]$$

$$= \frac{1}{N}\Sigma f\ cd \quad \left[u - \bar{u}\right]\left[v - \bar{v}\right]$$

$$= \frac{1}{N} cd\,\Sigma \quad f\,(u - \bar{u})\,(v - \bar{v})$$

$$= cd\ \frac{\Sigma f\,(u - \bar{u})\,(v - \bar{v})}{N} = cd \quad \text{cov}(u, v)$$

$\therefore \text{cov}(x, y) = c.d\ \text{cov}(u, v)$

$\therefore\ r_{xy} = \frac{\text{cov}(x, y)}{\sigma_x\ \sigma_y} = \frac{cd \quad \text{cov}(u, v)}{c \,.\sigma_u\, .\, d\,.\sigma_v} = \frac{\text{cov}(u, v)}{\sigma_u \quad \sigma_v} = r$

$\therefore r_{xy} = r_{uv}$

Steps:

1. Take the step deviations of the variable x and denote these deviations by u.
2. Take the step deviations of the variable y and denote these deviations by v.
3. Multiply uv and the respective frequency of each cell and unite the figure obtained in the right hand bottom corner of each cell.
4. Add the corrected (all) as calculated in step 3 and obtain the total Σfuv.
5. Multiply the frequencies of the variable x by the deviations of x and obtain the total Σfu.
6. Take the squares of the step deviations of the variable x and multiply them by the respective frequencies and obtain the Σfu^2

Similarly get Σfv and Σfv^2 . Then substitute these values in the formula 1 and get the value of ' r' .

Example 4:

The following are the marks obtained by 132 students in two tests.

Test-1 Test-2	30-40	40-50	50-60	60-70	70-80	Total
20-30	2	5	3			10
30-40	1	8	12	6		27
40-50		5	22	14	1	42
50-60		2	16	9	2	29
60-70		1	8	6	1	16
70-80			2	4	2	8
Total	3	21	63	39	6	132

Calculate the correlation coefficient.

Let x denote Test 1 marks.

Let y denote Test 2 marks.

$$u = \frac{x - 55}{10} \qquad v = \frac{y - 45}{10}$$

mid x / mid y	35	45	55	65	75	f	v	fv	fv^2	fuv
25	4 2 8	2 5 10	0 3 0	-	-	10	-2	-20	40	18
35	2 1 2	1 8 8	0 12 0	-1 6 -6	-	27	-1	-27	27	4
45		0 5 0	0 22 0	0 14 0	0 1 0	42	0	0	0	0
55		-1 2 -2	0 16 0	1 9 9	2 2 4	29	1	29	29	11
65		-2 1 -2	0 8 0	2 6 12	4 1 4	16	2	32	64	14
75			0 2 0	3 4 12	6 2 12	8	3	24	72	24
f	3	21	63	39	6	132	3	38	232	71
u	-2	-1	0	1	2	0				
fu	-6	-21	0	39	12	24				
fu^2	12	21	0	39	24	96				
fuv	10	14	0	27	20	71				

Check

$$r = \frac{N\,\Sigma fuv - (\Sigma fu)\,(\Sigma fv)}{\sqrt{[\,N\Sigma fu^2 - (\Sigma fu)^2].[\,N\Sigma fv^2 - (\Sigma fv)^2]}}$$

$$= \frac{132 \times 71 - 24 \times 38}{\sqrt{[132 \times 96 - (24)^2]\ [132 \times 232 - (38)^2]}}$$

$$= \frac{9372 - 912}{\sqrt{(12672 - 576)\ (30624 - 1444)}}$$

$$= \frac{8460}{109.96 \times 170.82} = \frac{8460}{18786.78} = 0.4503$$

Example 5:

Calculate Karl Pearson' s coefficient of correlation from the data given below:

Age in years

Marks	18	19	20	21	22
0- 5	-	-	-	3	1
5- 10	-	-	-	3	2
10-15	-	-	7	10	-
15-20	-	5	4	-	-
20-25	3	2	-	-	-

$$u = \frac{x - 12.5}{5}$$

$$v = \frac{y - 20}{1}$$

y / mid x	18	19	20	21	2	f	v	fv	fv^2	Fuv
2.5	-	-	-	-2 3 | -6	-4 1 -4	4	-2	-8	16	-10
7.5	-	-	-	-1 3 | -3	-2 2 | -4	5	-1	-5	5	-7
12.5	-	-	0 7 | 0	0 10 | 0	-	17	0	0	0	0
17.5	-	-1 5 | -5	0 4 | 0	-	-	9	1	9	9	-5
22.5	-4 3 | -12	-2 2 -4	-	-	-	5	2	10	20	-16
f	3	7	11	16	3	40	0	6	50	-38
u	-2	-1	0	1	2	0				
fu	-6	-7	0	16	6	9				
fu^2	12	7	0	16	12	47				
fuv	-12	-9	0	-9	-8	-38				

Check

$$r = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{\sqrt{[N\Sigma fu^2 - (\Sigma fu)^2].[N\Sigma fv^2 - (\Sigma fv)^2]}}$$

$$= \frac{40(-38) - 6 \times 9}{\sqrt{[40 \times 50 - 6^2].[40 \times 47 - 9^2]}}$$

$$= \frac{-1520 - 54}{\sqrt{(2000 - 36) \times (1880 - 81)}} = \frac{-1574}{\sqrt{1964 \times 1799}} = -0.8373$$

Properties of Correlation:

1. Correlation coefficient lies between –1 and +1

(i.e) $-1 \le r \le +1$

Let $x' = \frac{x - \bar{x}}{\sigma_x}$; $y' = \frac{y - \bar{y}}{\sigma_y}$

Since $\Sigma(x' + y')^2$ being sum of squares is always non-negative.

$\Sigma(x' + y')^2 \ge 0$

$\Sigma x'^2 + \Sigma y'^2 + 2\Sigma x' y' \ge 0$

$$\Sigma\left(\frac{x - \bar{x}}{\sigma_x}\right)^2 + \Sigma\left(\frac{y - \bar{y}}{\sigma_y}\right)^2 + 2\Sigma\left(\frac{x - \bar{x}}{\sigma_x}\right)\left(\frac{y - \bar{y}}{\sigma_y}\right) \ge 0$$

$$\frac{\Sigma(x - \bar{x})^2}{\sigma_{x^2}} + \frac{\Sigma(y - \bar{y})^2}{\sigma_{y^2}} + \frac{2\Sigma(x - \bar{x})(Y - \bar{Y})}{\sigma_x \sigma_y} \ge 0$$

dividing by ‘ n’ we get

$$\frac{1}{\sigma_{x^2}} \cdot \frac{1}{n} \Sigma(x - \bar{x})^2 + \frac{1}{\sigma_{y^2}} \cdot \frac{1}{n} \Sigma(y - \bar{y})^2 + \frac{2}{\sigma_x \sigma_y} \cdot \frac{1}{n} \Sigma(x - \bar{x})(y - \bar{y}) \ge 0$$

$$\frac{1}{\sigma_{x^2}} \cdot \sigma_{x^2} + \frac{1}{\sigma_{y^2}} \sigma_{y^2} + \frac{2}{\sigma_x \sigma_y} \cdot \mathrm{cov}(x, y) \ge 0$$

$1 + 1 + 2r \ge 0$

$2 + 2r \ge 0$

$2(1+r) \ge 0$

$(1 + r) \ge 0$

$-1 \leq r$ .....(1)

Similarly, $\Sigma(x' - y')^2 \geq 0$

$2(1-r) \geq 0$

$1 - r \geq 0$

$r \leq +1$ .....(2)

(1) +(2) gives $-1 \leq r \leq 1$

Note: r = +1 perfect +ve correlation.

r = –1 perfect –ve correlation between the variables.

Property 2: ' r' is independent of change of origin and scale.

Property 3: It is a pure number independent of units of measurement.

Property 4: Independent variables are uncorrelated but the converse is not true.

Property 5: Correlation coefficient is the geometric mean of two regression coefficients.

Property 6: The correlation coefficient of x and y is symmetric. $r_{xy} = r_{yx}$.

Limitations:

1. Correlation coefficient assumes linear relationship regardless of the assumption is correct or not.
2. Extreme items of variables are being unduly operated on correlation coefficient.
3. Existence of correlation does not necessarily indicate cause-effect relation.

Interpretation:

The following rules helps in interpreting the value of ' r' .

1. When r = 1, there is perfect +ve relationship between the variables.
2. When r = -1, there is perfect –ve relationship between the variables.
3. When r = 0, there is no relationship between the variables.
4. If the correlation is +1 or –1, it signifies that there is a high degree of correlation. (+ve or –ve) between the two variables.

If r is near to zero (ie) 0.1,-0.1, (or) 0.2 there is less correlation.

Rank Correlation:

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc.The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by Edward Spearman in 1904. It is defined as $r = 1 - \frac{6\Sigma D^2}{n^3 - n}$ r = rank correlation coefficient.

Note: Some authors use the symbol ρ for rank correlation.

ΣD^2 = sum of squares of differences between the pairs of ranks.

n = number of pairs of observations.

The value of r lies between –1 and +1. If r = +1, there is complete agreement in order of ranks and the direction of ranks is also same. If r = -1, then there is complete disagreement in order of ranks and they are in opposite directions.

Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an average rank is to be given to each individual item. For example if the value so is repeated twice at the 5th rank, the common rank to be assigned to each item is $\frac{5+6}{2}$ = 5.5 which is the average of 5 and 6 given as 5.5, appeared twice.

If the ranks are tied, it is required to apply a correction factor which is $\frac{1}{12}$ (m^3-m). A slightly different formula is used when there is more than one item having the same value.

The formula is

$$r = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots]}{n^3 - n}$$

Where m is the number of items whose ranks are common and should be repeated as many times as there are tied observations.

Example 6:

In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between and tea and coffee price.

Price of tea	88	90	95	70	60	75	50
Price of coffee	120	134	150	115	110	140	100

Price of tea	Rank	Price of coffee	Rank	D	D^2
88	3	120	4	1	1
90	2	134	3	1	1
95	1	150	1	0	0
70	5	115	5	0	0
60	6	110	6	0	0
75	4	140	2	2	4
50	7	100	7	0	0
					$\Sigma D^2 = 6$

$$r = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 6}{7^3 - 7}$$

$$= 1 - \frac{36}{336} = 1 - 0.1071$$

$$= 0.8929$$

The relation between price of tea and coffee is positive at 0.89. Based on quality the association between price of tea and price of coffee is highly positive.

Example 7:

In an evaluation of answer script the following marks are awarded by the examiners.

1^{st}	88	95	70	960	50	80	75	85
2^{nd}	84	90	88	55	48	85	82	72

Do you agree the evaluation by the two examiners is fair?

x	R1	y	R2	D	D^2
88	2	84	4	2	4
95	1	90	1	0	0
70	6	88	2	4	16
60	7	55	7	0	0
50	8	48	8	0	0
80	4	85	3	1	1
85	3	75	6	3	9
					30

$$r = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 30}{8^3 - 8}$$

$$= 1 - \frac{180}{504} = 1 - 0.357 = 0.643$$

r = 0.643 shows fair in awarding marks in the sense that uniformity has arisen in evaluating the answer scripts between the two examiners.

Example 8:

Rank Correlation for tied observations. Following are the marks obtained by 10 students in a class in two tests.

Students	A	B	C	D	E	F	G	H	I	J
Test 1	70	68	67	55	60	60	75	63	60	72
Test 2	65	65	80	60	68	58	75	63	60	70

Calculate the rank correlation coefficient between the marks of two tests.

Student	Test 1	R1	Test 2	R2	D	D^2
A	70	3	65	5.5	-2.5	6.25
B	68	4	65	5.5	-1.5	2.25
C	67	5	80	1.0	4.0	16.00
D	55	10	60	8.5	1.5	2.25
E	60	8	68	4.0	4.0	16.00
F	60	8	58	10.0	-2.0	4.00
G	75	1	75	2.0	-1.0	1.00
H	63	6	62	7.0	-1.0	1.00
I	60	8	60	8.5	0.5	0.25
J	72	2	70	3.0	-1.0	1.00
						50.00

60 is repeated 3 times in test 1.
60,65 is repeated twice in test 2.
m = 3; m = 2; m = 2

$$r = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)}{n^3 - n}$$

$$= 1 - \frac{6[50 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{10^3 - 10}$$

$$= 1 - \frac{6[50 + 2 + 0.5 + 0.5]}{990}$$

$$= 1 - \frac{6 \times 53}{990} = \frac{672}{990} = 0.68$$

Interpretation: There is uniformity in the performance of students in the two tests.

Exercise – 8

I. Choose the correct answer:

1. Limits for correlation coefficient.
   (a) $-1 \le r \le 1$ (b) $0 \le r \le 1$
   (c) $-1 \le r \le 0$ (d) $1 \le r \le 2$
2. The coefficient of correlation.
   (a) cannot be negative (b) cannot be positive
   (c) always positive (d)can either be positive or negative
3. The product moment correlation coefficient is obtained by
   (a) $r = \frac{\Sigma XY}{xy}$ (b) $r = \frac{\Sigma XY}{n \sigma_x \sigma_y}$
   (c) $r = \frac{\Sigma XY}{n \sigma_x}$ (d) none of these
4. If cov(x,y) = 0 then
   (a) x and y are correlated (b) x and y are uncorrelated
   (c) none (d) x and y are linearly related

5. If r = 0 the cov (x,y) is
   (a) 0 (b) -1 (c) 1 (d) 0.2
6. Rank correlation coefficient is given by
   (a) $1+\frac{6\Sigma D^2}{n^3-n}$ (b) $1-\frac{6\Sigma D^2}{n^2-n}$ (c) $1-\frac{6\Sigma D^2}{n^3-n}$
   (d) $1-\frac{6\Sigma D^2}{n^3+n}$
7. If cov (x,y) = $\sigma_x \sigma_y$ then
   (a) r = +1 (b) r = 0 (c) r = 2 (d) r = -1
8. If $\Sigma D^2 = 0$ rank correlation is
   (a) 0 (b) 1 (c)0.5 (d) -1
9. Correlation coefficient is independent of change of
   (a) Origin (b) Scale
   (c) Origin and Scale (d) None
10. Rank Correlation was found by
   (a) Pearson (b) Spearman
   (c) Galton (d) Fisher

II. Fill in the blanks:

11 Correlation coefficient is free from_____.
12 The diagrammatic representation of two variables is called _____
13 The relationship between three or more variables is studied with the help of_____correlation.
14 Product moment correlation was found by _____
15 When r = +1, there is_____correlation.
16 If $r_{xy} = r_{yx}$, correlation between x and y is _____
17 Rank Correlation is useful to study_____characteristics.
18 The nature of correlation for shoe size and IQ is _____

III. Answer the following :

19 What is correlation?
20 Distinguish between positive and negative correlation.
21 Define Karl Pearson' s coefficient of correlation. Interpret *r*, when *r* = 1, -1 and 0.
22 What is a scatter diagram? How is it useful in the study of Correlation?

23 Distinguish between linear and non-linear correlation.
24 Mention important properties of correlation coefficient.
25 Prove that correlation coefficient lies between –1 and +1.
26 Show that correlation coefficient is independent of change of origin and scale.
27 What is Rank correlation? What are its merits and demerits?
28 Explain different types of correlation with examples.
29 Distinguish between Karl Pearson' s coefficient of correlation and Spearman' s correlation coefficient.
30 For 10 observations $\Sigma x = 130$; $\Sigma y = 220$; $\Sigma x^2 = 2290$; $\Sigma y^2 = 5510$; $\Sigma xy = 3467$. Find ' r'.
31 Cov (x,y) = 18.6; var(x) = 20.2; var(y) = 23.7. Find ' r'.
32 Given that $r = 0.42$ cov(x,y) = 10.5 v(x) = 16; Find the standard deviation of y.
33 Rank correlation coefficient $r = 0.8.\Sigma D^2 = 33$. Find ' n' .

Karl Pearson Correlation:

34. Compute the coefficient of correlation of the following score of A and B.

A	5	10	5	11	12	4	3	2	7	1
B	1	6	2	8	5	1	4	6	5	2

35. Calculate coefficient of Correlation between price and supply. Interpret the value of correlation coefficient.

Price	8	10	15	17	20	22	24	25
Supply	25	30	32	35	37	40	42	45

36. Find out Karl Pearson' s coefficient of correlation in the following series relating to prices and supply of a commodity.

Price(Rs.)	11	12	13	14	15	16	17	18	19	20
Supply(Rs.)	30	29	29	25	24	24	24	21	18	15

37. Find the correlation coefficient between the marks obtained by ten students in economics and statistics.

Marks (in economics	70	68	67	55	60	60	75	63	60	72
Marks (in statistics	65	65	80	60	68	58	75	62	60	70

38. Compute the coefficient of correlation from the following data.

Age of workers	40	34	22	28	36	32	24	46	26	30
Days absent	2.5	3	5	4	2.5	3	4.5	2.5	4	3.5

39. Find out correlation coefficient between height of father and son from the following data

Height of father	65	66	67	67	68	69	70	72
Height of son	67	68	65	68	72	72	69	71

BI-VARIATE CORRELATION:

40. Calculate Karl Pearson' s coefficient of correlation.for the following data.

Class Interval	0	1	2	3	4	5	6	7	8	Total
20-29	2	1	2	2	-	1	-	1	1	10
30-39	-	2	-	1	-	2	-	1	2	8
40-49	-	2	-	2	-	-	1	-	1	6
50-59	1	-	2	-	-	-	-	1	-	4
60-69	-	-	-	-	-	1	-	1	-	2

41. Calculate the coefficient of correlation and comment upon your result.

Age of wives

Age of Husband	15-25	25-35	35-45	45-55	55-65	65-75	Total
15-25	1	1	-	-	-	-	2
25-35	2	12	1	-	-	-	15
35-45	-	4	10	1	-	-	15
45-55	-	-	3	6	1	-	10
55-65	-	-	-	2	4	2	8
65-75	-	-	-	-	1	2	3
Total	3	17	14	9	6	4	53

42. The following table gives class frequency distribution of 45 clerks in a business office according to age and pay. Find correlation between age and pay if any.

Pay

Age	60-70	70-80	80-90	90-100	100-110	Total
20-30	4	3	1	-	-	8
30-40	2	5	2	1	-	10
40-50	1	2	3	2	1	9
50-60	-	1	3	5	2	11
60-70	-	-	1	1	5	7
Total	7	11	10	9	8	45

43. Find the correlation coefficient between two subjects marks scored by 60 candidates.

Marks in Statistics

Marks in economics	5-15	15-25	25-35	35-45	Total
0-10	1	1	-	-	2
10-20	3	6	5	1	15
20-30	1	8	9	2	20
30-40	-	3	9	3	15
40-50	-	-	4	4	8
Total	5	18	27	10	60

44. Compute the correlation coefficient for the following data.

Advertisement Expenditure(‘ 000)

Sales Revenue (Rs.’ 000)	5-15	15-25	25-35	35-45	Total
75-125	4	1	-	-	5
125-175	7	6	2	1	16
175-225	1	3	4	2	10
225-275	1	1	3	4	9
Total	13	11	9	7	40

45. The following table gives the no. of students having different heights and weights. Do you find any relation between height and weight.

Weights in Kg

Height in cms	55-60	60-65	65-70	70-75	75-80	Total
150-155	1	3	7	5	2	18
155-160	2	4	10	7	4	27
160-165	1	5	12	10	7	35
165-170	-	3	8	6	3	20
Total	4	15	37	28	16	100

RANK CORRELATION:

46. Two judges gave the following ranks to eight competitors in a beauty contest. Examine the relationship between their judgements.

Judge A	4	5	1	2	3	6	7	8
Judge B	8	6	2	3	1	4	5	7

47. From the following data, calculate the coefficient of rank correlation.

X	36	56	20	65	42	33	44	50	15	60
Y	50	35	70	25	58	75	60	45	80	38

48. Calculate spearman' s coefficient of Rank correlation for the following data.

X	53	98	95	81	75	71	59	55
Y	47	25	32	37	30	40	39	45

49. Apply spearman' s Rank difference method and calculate coefficient of correlation between x and y from the data given below.

X	22	28	31	23	29	31	27	22	31	18
Y	18	25	25	37	31	35	31	29	18	20

50. Find the rank correlation coefficients.

Marks in Test I	70	68	67	55	60	60	75	63	60	72
Marks in Test II	65	65	80	60	68	58	75	62	60	70

51. Calculate spearman' s Rank correlation coefficient for the following table of marks of students in two subjects.

First subject	80	64	54	49	48	35	32	29	20	18	15	10
Second subject	36	38	39	41	27	43	45	52	51	42	40	52

IV. Suggested Activities

Select any ten students from your class and find their heights and weights. Find the correlation between their heights and weights

Answers:

I.

1. (a). 2. (d) 3. (b) 4.(b) 5. (a)
6. (c) 7. (a) 8. (b) 9. (c) 10. (b)

II.

11. Units 12. Scatter diagram 13. Multiple
14. Pearson 15. Positive perfect 16. Symmetric
17. Qualitative 18. No correlation

III.

30. $r = 0.9574$ 31. $r = 0.85$ 32. $o_y = 6.25$.
33. n = 10 34. $r = +0.58$ 35. $r = +0.98$
36. $r = - 0.96$ 37. $r = +0.68$ 38. $r = - 0.92$
39. $r = +0.64$ 40. $r = +0.1$ 41. $r = +0.98$
42. $r = +0.746$ 43. $r = +0.533$ 44. $r = +0.596$
45. $r = +0.0945$ 46. $r = +0.62$ 47. $r = - 0.93$
48. $r = - 0.905$ 49. $r = 0.34$ 50. $r = 0.679$
51. $r = 0.685$

9. REGRESSION

Introduction:

After knowing the relationship between two variables we may be interested in estimating (predicting) the value of one variable given the value of another. The variable predicted on the basis of other variables is called the "dependent" or the ' explained' variable and the other the ' independent' or the ' predicting' variable. The prediction is based on average relationship derived statistically by regression analysis. The equation, linear or otherwise, is called the regression equation or the explaining equation.

For example, if we know that advertising and sales are correlated we may find out expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales.

The relationship between two variables can be considered between, say, rainfall and agricultural production, price of an input and the overall cost of product, consumer expenditure and disposable income. Thus, regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

Definition:

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

Types Of Regression:

The regression analysis can be classified into:

a) Simple and Multiple
b) Linear and Non –Linear
c) Total and Partial

a) Simple and Multiple:

In case of simple relationship only two variables are considered, for example, the influence of advertising expenditure on sales turnover. In the case of multiple relationship, more than

two variables are involved. On this while one variable is a dependent variable the remaining variables are independent ones.

For example, the turnover (y) may depend on advertising expenditure (x) and the income of the people (z). Then the functional relationship can be expressed as y = f (x,z).

b) Linear and Non-linear:

The linear relationships are based on straight-line trend, the equation of which has no-power higher than one. But, remember a linear relationship can be both simple and multiple. Normally a linear relationship is taken into account because besides its simplicity, it has a better predective value, a linear trend can be easily projected into the future. In the case of non-linear relationship curved trend lines are derived. The equations of these are parabolic.

c) Total and Partial:

In the case of total relationships all the important variables are considered. Normally, they take the form of a multiple relationships because most economic and business phenomena are affected by multiplicity of cases. In the case of partial relationship one or more variables are considered, but not all, thus excluding the influence of those not found relevant for a given purpose.

Linear Regression Equation:

If two variables have linear relationship then as the independent variable (X) changes, the dependent variable (Y) also changes. If the different values of X and Y are plotted, then the two straight lines of best fit can be made to pass through the plotted points. These two lines are known as regression lines. Again, these regression lines are based on two equations known as regression equations. These equations show best estimate of one variable for the known value of the other. The equations are linear.

Linear regression equation of Y on X is

$$Y = a + bX \ldots\ldots (1)$$

And X on Y is

$$X = a + bY \ldots\ldots (2)$$

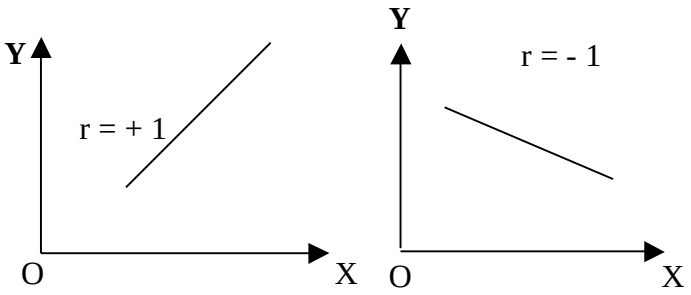
a, b are constants.

From (1) We can estimate Y for known value of X.
(2) We can estimate X for known value of Y.

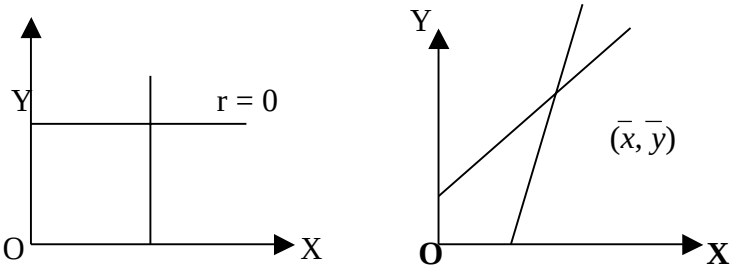
Regression Lines:

For regression analysis of two variables there are two regression lines, namely Y on X and X on Y. The two regression lines show the average relationship between the two variables.

For perfect correlation, positive or negative i.e., $r = \pm 1$, the two lines coincide i.e., we will find only one straight line. If $r = 0$, i.e., both the variables are independent then the two lines will cut each other at right angle. In this case the two lines will be parallel to X and Y-axes.



Lastly the two lines intersect at the point of means of X and Y. From this point of intersection, if a straight line is drawn on X-axis, it will touch at the mean value of x. Similarly, a perpendicular drawn from the point of intersection of two regression lines on Y-axis will touch the mean value of Y.



Principle of ' Least Squares' :

Regression shows an average relationship between two variables, which is expressed by a line of regression drawn by the method of "least squares". This line of regression can be derived graphically or algebraically. Before we discuss the various methods let us understand the meaning of least squares.

A line fitted by the method of least squares is known as the line of best fit. The line adapts to the following rules:

(i) The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero. i.e.,

$\sum(X - Xc) = 0$ or $\sum (Y - Yc) = 0$

Where Xc and Yc are the values obtained by regression analysis.

(ii) The sum of the squares of these deviations is less than the sum of squares of deviations from any other line. i.e.,

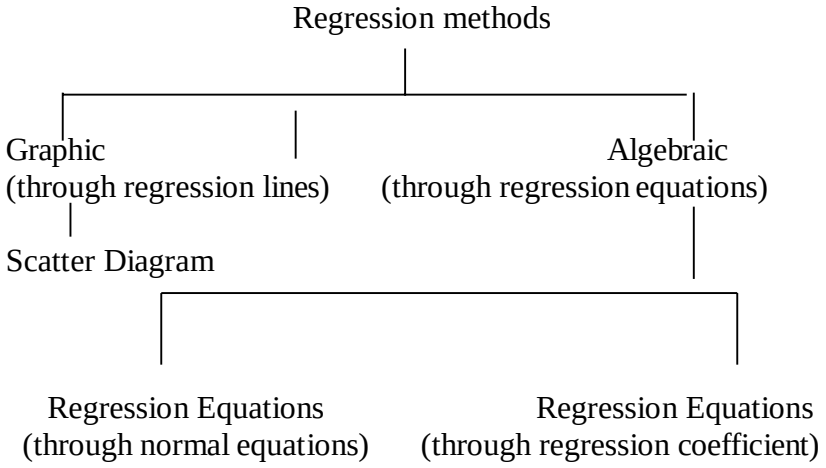
$\sum(Y - Yc)^2 < \sum (Y - Ai)^2$

Where Ai = corresponding values of any other straight line.

(iii) The lines of regression (best fit) intersect at the mean values of the variables X and Y, i.e., intersecting point is $\bar{x}, \bar{y}$.

Methods of Regression Analysis:

The various methods can be represented in the form of chart given below:



Graphic Method:

Scatter Diagram:

Under this method the points are plotted on a graph paper representing various parts of values of the concerned variables. These points give a picture of a scatter diagram with several points spread over. A regression line may be drawn in between these points either by free hand or by a scale rule in such a way that the squares of the vertical or the horizontal distances (as the case may be) between the points and the line of regression so drawn is the least. In other words, it should be drawn faithfully as the line of best fit leaving equal number of points on both sides in such a manner that the sum of the squares of the distances is the best.

Algebraic Methods:

(i) Regression Equation.

The two regression equations

for X on Y; $X = a + bY$

And for Y on X; $Y = a + bX$

Where X, Y are variables, and a,b are constants whose values are to be determined

For the equation, $X = a + bY$

The normal equations are

$$\sum X = na + b \sum Y \text{ and}$$

$$\sum XY = a\sum Y + b\sum Y^2$$

For the equation, $Y = a + bX$, the normal equations are

$$\sum Y = na + b\sum X \text{ and}$$

$$\sum XY = a\sum X + b\sum X^2$$

From these normal equations the values of a and b can be determined.

Example 1:

Find the two regression equations from the following data:

X:	6	2	10	4	8
Y:	9	11	5	8	7

Solution:

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
30	40	220	340	214

Regression equation of Y on X is $Y = a + bX$ and the normal equations are

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values, we get

$$40 = 5a + 30b \ldots\ldots (1)$$
$$214 = 30a + 220b \ldots\ldots (2)$$

Multiplying (1) by 6

$$240 = 30a + 180b \ldots\ldots (3)$$

(2) – (3) $\quad -26 = 40b$

$$\text{or } b = -\frac{26}{40} = -0.65$$

Now, substituting the value of 'b' in equation (1)

$$40 = 5a - 19.5$$
$$5a = 59.5$$
$$a = \frac{59.5}{5} = 11.9$$

Hence, required regression line Y on X is $Y = 11.9 - 0.65\,X$.

Again, regression equation of X on Y is

$X = a + bY$ and

The normal equations are

$$\sum X = na + b\sum Y \text{ and}$$
$$\sum XY = a\sum Y + b\sum Y^2$$

Now, substituting the corresponding values from the above table, we get

$30 = 5a + 40b$ ….(3)

$214 = 40a + 340b$ ….(4)

Multiplying (3) by 8, we get

$240 = 40a + 320\ b$ ….(5)

(4) – (5) gives

$-26 = 20b$

$b = -\frac{26}{20} = -1.3$

Substituting $b = -1.3$ in equation (3) gives

$30 = 5a - 52$

$5a = 82$

$a = \frac{82}{5} = 16.4$

Hence, Required regression line of X on Y is

X = 16.4 – 1.3Y

(ii) Regression Co-efficents:

The regression equation of Y on X is $y_e = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

Here, the regression Co.efficient of Y on X is

$b_1 = b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$y_e = \bar{y} + b_1(x - \bar{x})$

The regression equation of X on Y is

$X_e = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

Here, the regression Co-efficient of X on Y

$b_2 = b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$X_e = \bar{X} + b_2(y - \bar{y})$

If the deviation are taken from respective means of x and y

$$b_1 = b_{yx} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{\sum xy}{\sum x^2} \quad \text{and}$$

$$b_2 = b_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2} = \frac{\sum xy}{\sum y^2}$$

where $x = X - \bar{X}$, $y = Y - \bar{Y}$

If the deviations are taken from any arbitrary values of x and y (short – cut method)

$$b_1 = b_{yx} = \frac{n\sum uv - \sum u \sum v}{n\sum u^2 - \left(\sum u\right)^2}$$

$$b_2 = b_{xy} = \frac{n\sum uv - \sum u \sum v}{n\sum v^2 - \left(\sum v\right)^2}$$

where u = x – A : v = Y-B

A = any value in X

B = any value in Y

Properties of Regression Co-efficient:

1. Both regression coefficients must have the same sign, ie either they will be positive or negative.
2. correlation coefficient is the geometric mean of the regression coefficients ie, $r = \pm\sqrt{b_1 b_2}$
3. The correlation coefficient will have the same sign as that of the regression coefficients.
4. If one regression coefficient is greater than unity, then other regression coefficient must be less than unity.
5. Regression coefficients are independent of origin but not of scale.
6. Arithmetic mean of b_1 and b_2 is equal to or greater than the coefficient of correlation. Symbolically $\frac{b_1 + b_2}{2} \geq r$

7. If r=0, the variables are uncorrelated , the lines of regression become perpendicular to each other.
8. If r= $\pm$1, the two lines of regression either coincide or parallel to each other
9. Angle between the two regression lines is $\theta = \tan^{-1}\left[\frac{m_1 - m_2}{1 + m_1 m_2}\right]$

where m_1 and,m_2 are the slopes of the regression lines X on Y and Y on X respectively.

10.The angle between the regression lines indicates the degree of dependence between the variables.

Example 2:

If 2 regression coefficients are $b_1 = \frac{4}{5}$ and $b_2 = \frac{9}{20}$ .What would be the value of r?

Solution:

The correlation coefficient , $r = \pm\sqrt{b_1 b_2}$

$$= \sqrt{\frac{4}{5} \times \frac{9}{20}}$$

$$= \sqrt{\frac{36}{100}} = \frac{6}{10} = 0.6$$

Example 3:

Given $b_1 = \frac{15}{8}$ and $b_2 = \frac{3}{5}$, Find r

Solution:

$r = \pm\sqrt{b_1 b_2}$

$= \sqrt{\frac{15}{8} \times \frac{3}{5}}$

$= \sqrt{\frac{9}{8}} = 1.06$

It is not possible since *r*, cannot be greater than one. So the given values are wrong

9.6 Why there are two regression equations?

The regression equation of Y on X is

$$Y_e = \bar{Y} + r\frac{\sigma_y}{\sigma_x}(X - \bar{X}) \quad \text{(or)} \quad Y_e = \bar{Y} + b_1(X - \bar{X}) \qquad (1)$$

The regression equation of X on Y is

$$X_e = \bar{X} + r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$X_e = \bar{X} + b_2(Y - \bar{Y})$$

These two regression equations represent entirely two different lines. In other words, equation (1) is a function of X, which can be written as $Y_e = F(X)$ and equation (2) is a function of Y, which can be written as $X_e = F(Y)$.

The variables X and Y are not inter changeable. It is mainly due to the fact that in equation (1) Y is the dependent variable, X is the independent variable. That is to say for the given values of X we can find the estimates of Ye of Y only from equation (1). Similarly, the estimates Xe of X for the values of Y can be obtained only from equation (2).

Example 4:

Compute the two regression equations from the following data.

X	1	2	3	4	5
Y	2	3	5	4	6

If x =2.5, what will be the value of y?

Solution:

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
1	2	-2	-2	4	4	4
2	3	-1	-1	1	1	-1
3	5	0	1	0	1	0
4	4	1	0	1	0	0
5	6	2	2	4	4	4
15	20	20		10	10	9

$$\overline{X} = \frac{\sum X}{n} = \frac{15}{5} = 3$$

$$\overline{Y} = \frac{\sum Y}{n} = \frac{20}{5} = 4$$

Regression Co efficient of Y on X

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{9}{10} = 0.9$$

Hence regression equation of Y on X is

$$Y = \overline{Y} + b_{yx}(X - \overline{X})$$

$= 4 + 0.9 (X – 3)$

$= 4 + 0.9X – 2.7$

$=1.3 + 0.9X$

when X = 2.5

$Y = 1.3 + 0.9 \times 2.5$

$= 3.55$

Regression co efficient of X on Y

$$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{9}{10} = 0.9$$

So, regression equation of X on Y is

$$X = \overline{X} + b_{xy}(Y - \overline{Y})$$

$= 3 + 0.9 (Y – 4)$

$= 3 + 0.9Y – 3.6$

$= 0.9Y - 0.6$

Short-cut method

Example 5:

Obtain the equations of the two lines of regression for the data given below:

X	45	42	44	43	41	45	43	40
Y	40	38	36	35	38	39	37	41

Solution:

X	Y	u = X-A	u^2	v = Y-B	V^2	uv
46	40	3	9	2	4	6
42	38 **B**	-1	1	0	0	0
44	36	1	1	-2	4	-2
A 43	35	0	0	-3	9	0
41	38	-2	4	0	0	0
45	39	2	4	1	1	2
43	37	0	0	-1	1	0
40	41	-3	9	3	9	-9
		0	28	0	28	-3

$$X = A + \frac{\sum u}{n}$$

$$= 43 + \frac{0}{8} \quad = 43$$

$$Y = B + \frac{\sum u}{n}$$

$$= 38 + \frac{0}{8} \quad = 38$$

The regression Co-efficient of Y on X is

$$b_1 = b_{yx} = \frac{n\sum uv - \sum u \sum v}{n\sum u^2 - \left(\sum u\right)^2}$$

$$= \frac{8(-3) - (0)(0)}{8(28) - (0)^2} \quad = \frac{-24}{224} \quad = -0.11$$

The regression coefficient of X on Y is

$$b_2 = b_{xy} = \frac{n\sum uv - \sum u \sum v}{n\sum v^2 - \left(\sum v\right)^2}$$

$$= \frac{8(-3) - (0)(0)}{8(28) - (0)^2}$$

$$= \frac{-24}{224} \quad = -\ 0.11$$

Hence the reression equation of Y on X is

$$Y_e = \overline{Y} + b_1 (X - \overline{X})$$

= 38 – 0.11 (X-43)

= 38 – 0.11X + 4.73

= 42.73 – 0.11X

The regression equation of X on Y is

$$X_e = \overline{X} + b_1 (Y - \overline{Y})$$

= 43 – 0.11 (Y-38)

= 43 – 0.11Y + 4.18

= 47.18 – 0.11Y

Example 6:

In a correlation study, the following values are obtained

	X	Y
Mean	65	67
S.D	2.5	3.5

Co-efficient of correlation = 0.8

Find the two regression equations that are associated with the above values.

Solution:

Given,

$\overline{X}$ = 65, $\overline{Y}$ = 67, σ_x = 2.5, σ_y= 3.5, r = 0.8

The regression co-efficient of Y on X is

$$b_{yx} = b_1 = r \frac{\sigma_y}{\sigma_x}$$

$$= 0.8 \times \frac{3.5}{2.5} = 1.12$$

The regression coefficient of X on Y is

$$b_{xy} = b_2 = r \frac{\sigma_x}{\sigma_y}$$

$$= 0.8 \times \frac{2.5}{3.5} = 0.57$$

Hence, the regression equation of Y on X is

$$Y_e = \bar{Y} + b_1 (X - \bar{X})$$

$$= 67 + 1.12 \text{ (X-65)}$$

$$= 67 + 1.12 \text{ X} - 72.8$$

$$= 1.12\text{X} - 5.8$$

The regression equation of X on Y is

$$X_e = \bar{X} + b_2 (Y - \bar{Y})$$

$$= 65 + 0.57 \text{ (Y-67)}$$

$$= 65 + 0.57\text{Y} - 38.19$$

$$= 26.81 + 0.57\text{Y}$$

Note:

Suppose, we are given two regression equations and we have not been mentioned the regression equations of Y on X and X on Y. To identify, always assume that the first equation is Y on X then calculate the regression co-efficient $b_{yx} = b_1$ and $b_{xy} = b_2$. If these two are satisfied the properties of regression co-efficient, our assumption is correct, otherwise interchange these two equations.

Example 7:

Given 8X – 10Y + 66 = 0 and 40X – 18Y = 214. Find the correlation coefficient, *r*.

Solution:

Assume that the regression equation of Y on X is 8X- 10Y + 66 = 0.

$$-10\text{Y} = -66 - 8\text{X}$$

$$10\text{Y} = 66 + 8\text{X}$$

$$\text{Y} = \frac{66}{10} + \frac{8X}{10}$$

Now the coefficient attached with X is b_{yx}

i.e., $b_{yx} = \frac{8}{10} = \frac{4}{5}$

The regression equation of X on Y is

$40X-18Y=214$

In this keeping X left side and write other things right side

i.e., $40X = 214 + 18Y$

i.e., $X = \frac{214}{40} + \frac{18}{40}Y$

Now, the coefficient attached with Y is b_{xy}

i.e., $b_{xy} = \frac{18}{40} = \frac{9}{20}$

Here b_{yx} and b_{xy} are satisfied the properties of regression coefficients, so our assumption is correct.

Correlation Coefficient, $r = \sqrt{b_{yx} b_{xy}}$

$= \sqrt{\frac{4}{5} \times \frac{9}{20}}$

$= \sqrt{\frac{36}{100}}$

$= \frac{6}{10}$

$= 0.6$

Example 8:

Regression equations of two correlated variables X and Y are $5X-6Y+90 = 0$ and $15X-8Y-130 = 0$. Find correlation coefficient.

Solution:

Let $5X-6Y+90 =0$ represents the regression equation of X on Y and other for Y on X

Now $X= \frac{6}{5} Y - \frac{90}{5}$

$$b_{xy} = b_2 = \frac{6}{5}$$

For $15X-8Y-130 = 0$

$$Y = \frac{15}{8} X - \frac{130}{8}$$

$$b_{yx} = b_1$$

$$= \frac{15}{8}$$

$$r = \pm \sqrt{b_1 b_2}$$

$$= \sqrt{\frac{15}{8} \times \frac{6}{5}}$$

$$= \sqrt{2.25}$$

$$= 1.5 > 1$$

It is not possible. So our assumption is wrong. So let us take the first equation as Y on X and second equation as X on Y.

From the equation $5x - 6y + 90 = 0$,

$$Y = \frac{5}{6} X - \frac{90}{6}$$

$$b_{yx} = \frac{5}{6}$$

From the equation $15x - 8y - 130 = 0$,

$$X = \frac{8}{15} Y + \frac{130}{15}$$

$$b_{xy} = \frac{8}{15}$$

Correlation coefficient, $r = \pm \sqrt{b_1 b_2}$

$$= \sqrt{\frac{5}{6} \times \frac{8}{15}}$$

$$= \sqrt{\frac{40}{90}}$$

$$= \frac{2}{3}$$

$= 0.67$

Example 9:

The lines of regression of Y on X and X on Y are respectively, $y = x + 5$ and $16X = 9Y - 94$. Find the variance of X if the variance of Y is 19. Also find the covariance of X and Y.

Solution:

From regression line Y on X,

$Y = X+5$

We get $b_{yx} = 1$

From regression line X on Y,

$16X = 9Y-94$

$X = \frac{9}{16} Y - \frac{94}{16}$,

we get

$b_{xy} = \frac{9}{16}$

$r = \pm \sqrt{b_1 b_2}$

$= \sqrt{1 \times \frac{9}{16}}$

$= \frac{3}{4}$

Again , $b_{yx} = r \frac{\sigma_y}{\sigma_x}$

i.e., $1 = \frac{3}{4} \times \frac{4}{a_x}$ (Since $\sigma_Y{}^2 = 16$, $\sigma_Y = 4$)

$\sigma_X = 3$.

Variance of X $= \sigma_X{}^2$

$= 9$

Again $b_{yx} = \frac{\text{cov}(x, y)}{a_x{}^2}$

$$1 = \frac{\text{cov}(x, y)}{9}$$

or cov (x,y) = 9.

Example 10:

Is it possible for two regression lines to be as follows: Y = -1.5X + 7 , X = 0.6Y + 9 ? Give reasons.

Solution:

The regression coefficient of Y on X is $b_1 = b_{yx} = -1.5$
The regression coefficient of X on Y is $b_2 = b_{xy} = 0.6$
Both the regression coefficients are of different sign, which is a contrary. So the given equations cannot be regression lines.

Example 11:

In the estimation of regression equation of two variables X and Y the following results were obtained.

$\overline{X} = 90$, $\overline{Y} = 70$, n = 10, $\Sigma x^2 = 6360$; $\Sigma y^2 = 2860$, $\Sigma xy = 3900$ Obtain the two regression equations.

Solution:

Here, x, y are the deviations from the Arithmetic mean.

$$b_1 = b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

$$= \frac{3900}{6360} = 0.61$$

$$b_2 = b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$$

$$= \frac{3900}{2860} = 1.36$$

Regression equation of Y on X is

$Y_e = \overline{Y} + b_1 (X - \overline{X})$

$= 70 + 0.61 (X - 90)$

$= 70 + 0.61 X - 54.90$

$= 15.1 + 0.61X$

Regression equation of X on Y is

$X_e = \bar{X} + b_2 (Y - \bar{Y})$

$= 90 + 1.36 (Y - 70)$

$= 90 + 1.36 Y - 95.2 = 1.36Y - 5.2$

Uses of Regression Analysis:

1. Regression analysis helps in establishing a functional relationship between two or more variables.
2. Since most of the problems of economic analysis are based on cause and effect relationships, the regression analysis is a highly valuable tool in economic and business research.
3. Regression analysis predicts the values of dependent variables from the values of independent variables.
4. We can calculate coefficient of correlation (r) and coefficient of determination (r^2) with the help of regression coefficients.
5. In statistical analysis of demand curves, supply curves, production function, cost function, consumption function etc., regression analysis is widely used.

Difference between Correlation and Regression:

S.No	Correlation	Regression
1.	Correlation is the relationship between two or more variables, which vary in sympathy with the other in the same or the opposite direction.	Regression means going back and it is a mathematical measure showing the average relationship between two variables
2.	Both the variables X and Y are random variables	Here X is a random variable and Y is a fixed variable. Sometimes both the variables may be random variables.
3.	It finds out the degree of relationship between two variables and not the cause and effect of the variables.	It indicates the causes and effect relationship between the variables and establishes functional relationship.

4.	It is used for testing and verifying the relation between two variables and gives limited information.	Besides verification it is used for the prediction of one value, in relationship to the other given value.
5.	The coefficient of correlation is a relative measure. The range of relationship lies between –1 and +1	Regression coefficient is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable.
6.	There may be spurious correlation between two variables.	In regression there is no such spurious regression.
7.	It has limited application, because it is confined only to linear relationship between the variables.	It has wider application, as it studies linear and non-linear relationship between the variables.
8.	It is not very useful for further mathematical treatment.	It is widely used for further mathematical treatment.
9.	If the coefficient of correlation is positive, then the two variables are positively correlated and vice-versa.	The regression coefficient explains that the decrease in one variable is associated with the increase in the other variable.

Exercise – 9

I. Choose the correct answer:

1. When the correlation coefficient r = $\pm$1, then the two regression lines
   a) are perpendicular to each other b) coincide
   c) are parallel to each other d) none of these

2.If one regression coefficient is greater than unity then the other must be

a) greater than unity b) equal to unity
c) less than unity d) none of these

3. Regression equation is also named as

a) predication equation b) estimating equation
c) line of average relationship d) all the above

4. The lines of regression intersect at the point

a) (X,Y) b) $(\bar{X}, \bar{Y})$ c) (0,0) d) (1,1)

5. If r = 0, the lines of regression are

a) coincide b) perpendicular to each other
c) parallel to each other d) none of the above

6. Regression coefficient is independent of

a) origin b) scale c)both origin and scale
d) neither origin nor scale.

7. The geometric mean of the two-regression coefficients byx and bxy is equal to

a) r b) r^2 c) 1 d) $\sqrt{r}$

8. Given the two lines of regression as 3X – 4Y +8 = 0 and 4X – 3Y = 1, the means of X and Y are

a) X = 4, Y = 5 b) X =3, Y = 4
c) X = 2, Y = 2 d) X = 4/3, Y = 5/3

9. If the two lines of regression are
X + 2Y – 5 = 0 and
2X + 3Y – 8 = 0, the means of X and Y are

a) X = -3, Y = 4 b) X = 2, Y = 4
c) X =1, Y = 2 d) X = -1, Y = 2

10. If byx = -3/2, bxy = -3/2 then the correlation coefficient, r is

a) 3/2 b) –3/2 c) 9/4 d) – 9/4

II. Fill in the blanks:

11. The regression analysis measures _____ between X and Y.

12. The purpose of regression is to study_____between variables.

13. If one of the regression coefficients is_____unity, the other must be_____unity.

14. The farther the two regression lines cut each other, the _____ be the degree of correlation.
15. When one regression coefficient is positive, the other would also be_____.
16. The sign of regression coefficient is_____as that of correlation coefficient.

III. Answer the following:

17. Define regression and write down the two regression equations
18. Describe different types of regression.
19. Explain principle of least squares.
20. Explain (i) graphic method, (ii) Algebraic method.
21. What are regression co-efficient?
22. State the properties of regression coefficients.
23. Why there are two regression equations?
24. What are the uses of regression analysis?
25. Distinguish between correlation and regression.
26. What do you mean by regression line of Y on X and regression line of X on Y?
27. From the following data, find the regression equation
    $\Sigma X = 21$, $\Sigma Y = 20$, $\Sigma X^2 = 91$, $\Sigma XY = 74$, $n = 7$
28. From the following data find the regression equation of Y on X. If X = 15, find Y?

X	8	11	7	10	12	5	4	6
Y	11	30	25	44	38	25	20	27

29. Find the two regression equations from the following data.

X	25	22	28	26	35	20	22	40	20	18
Y	18	15	20	17	22	14	16	21	15	14

30. Find S.D (Y), given that variance of X = 36, bxy = 0.8, r = 0.5
31. In a correlation study, the following values are obtained

	X	Y
Mean	68	60
S.D.	2.5	3.5

Coefficient of correlation, r = 0.6 Find the two regression equations.

32. In a correlation studies, the following values are obtained:

	X	Y
Mean	12	15
S.D.	2	3

r = 0.5 Find the two regression equations.

33. The correlation coefficient of bivariate X and Y is r=0.6, variance of X and Y are respectively, 2.25 and 4.00, $\overline{X}$ =10, $\overline{Y}$ =20. From the above data, find the two regression lines

34. For the following lines of regression find the mean values of X and Y and the two regression coefficients

8X-10Y+66=0

40X-18Y=214

35. Given X=90, Y=70,bxy = 1.36, byx = 0.61

Find (i) the most probable values of X, when Y = 50 and

(ii) the coefficient of correlation between X and Y

36. You are supplied with the following data:

4X-5Y+33 = 0 and 20X-9Y-107 = 0

variance of Y = 4. Calculate

(I) Mean values of x and y

(II) S.D. of X

(III) Correlation coefficients between X and Y.

Answers:

I. 1. b 2. c 3.d 4. b 5. b 6. a 7. a 8. a 9. c 10. b

II.

11. dependence 12. dependence 13. more than, less than

14. lesser 15. positive 16. same.

III.

27. Y = 0.498X +1.366 28. Y =1.98X + 12.9;Y=42.6 30. 3.75

31. Y=2.88 + 0.84X, X = 42.2 + 0.43Y

32. Y = 6 + 0.75X ; X = 7 + 0.33Y

33. Y= 0.8X + 12, X = 0.45Y +1

34. $\overline{X}$ =13, $\overline{Y}$ = 17 b_{yx} = 9/20, b_{xy} = 4/5

35. (i) 62.8, (ii) 0.91

36. $\overline{X}$ =13, $\overline{Y}$ =17, S.D(X)=9, r = 0.6

10. INDEX NUMBERS

Introduction:

An index number is a statistical device for comparing the general level of magnitude of a group of related variables in two or more situation. If we want to compare the price level of 2000 with what it was in 1990, we shall have to consider a group of variables such as price of wheat, rice, vegetables, cloth, house rent etc., If the changes are in the same ratio and the same direction, we face no difficulty to find out the general price level. But practically, if we think changes in different variables are different and that too, upward or downward, then the price is quoted in different units i.e milk for litre, rice or wheat for kilogram, rent for square feet, etc

We want one figure to indicate the changes of different commodities as a whole. This is called an Index number. Index Number is a number which indicate the changes in magnitudes. M.Spiegel say, " An index number is a statistical measure designed to show changes in variable or a group of related variables with respect to time, geographic location or other characteristic". In general, index numbers are used to measure changes over time in magnitude which are not capable of direct measurement.

On the basis of study and analysis of the definition given above, the following characteristics of index numbers are apparent.

1. Index numbers are specified averages.
2. Index numbers are expressed in percentage.
3. Index numbers measure changes not capable of direct measurement.
4. Index numbers are for comparison.

Uses of Index numbers

Index numbers are indispensable tools of economic and business analysis. They are particular useful in measuring relative changes. Their uses can be appreciated by the following points.

1. They measure the relative change.
2. They are of better comparison.

3. They are good guides.
4. They are economic barometers.
5. They are the pulse of the economy.
6. They compare the wage adjuster.
7. They compare the standard of living.
8. They are a special type of averages.
9. They provide guidelines to policy.
10. To measure the purchasing power of money.

Types of Index numbers:

There are various types of index numbers, but in brief, we shall take three kinds and they are

(a) Price Index, (b) Quantity Index and (c) Value Index

(a) Price Index:

For measuring the value of money, in general, price index is used. It is an index number which compares the prices for a group of commodities at a certain time as at a place with prices of a base period. There are two price index numbers such as whole sale price index numbers and retail price index numbers. The wholesale price index reveals the changes into general price level of a country, but the retail price index reveals the changes in the retail price of commodities such as consumption of goods, bank deposits, etc.

(b) Quantity Index:

Quantity index number is the changes in the volume of goods produced or consumed. They are useful and helpful to study the output in an economy.

(c) Value Index

Value index numbers compare the total value of a certain period with total value in the base period. Here total value is equal to the price of commodity multiplied by the quantity consumed.

Notation: For any index number, two time periods are needed for comparison. These are called the Base period and the Current period. The period of the year which is used as a basis for comparison is called the base year and the other is the current year. The various notations used are as given below:

P_1 = Price of current year P_0 = Price of base year

q_1 = Quantity of current year q_0 = Quantity of base year

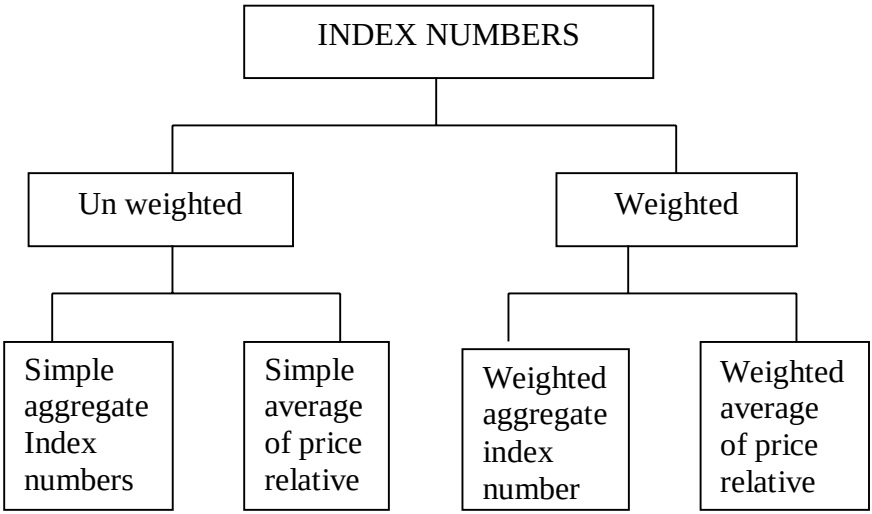
Problems in the construction of index numbers

No index number is an all purpose index number. Hence, there are many problems involved in the construction of index numbers, which are to be tackled by an economist or statistician. They are

1. Purpose of the index numbers
2. Selection of base period
3. Selection of items
4. Selection of source of data
5. Collection of data
6. Selection of average
7. System of weighting

Method of construction of index numbers:

Index numbers may be constructed by various methods as shown below:



10.5.1 Simple Aggregate Index Number

This is the simplest method of construction of index numbers. The price of the different commodities of the current year are added and the sum is divided by the sum of the prices of those commodities by 100. Symbolically,

$$\text{Simple aggregate price index} = P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

Where , Σp_1 = total prices for the current year

Σp_0 = Total prices for the base year

Example 1:

Calculate index numbers from the following data by simple aggregate method taking prices of 2000 as base.

Commodity	Price per unit (in Rupees)	
	2000	2004
A	80	95
B	50	60
C	90	100
D	30	45

Solution:

Commodity	Price per unit (in Rupees)	
	2000 (P_0)	2004 (P_1)
A	80	95
B	50	60
C	90	100
D	30	45
Total	250	300

Simple aggregate Price index = $P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$

$= \frac{300}{250} \times 100 = 120$

Simple Average Price Relative index:

In this method, first calculate the price relative for the various commodities and then average of these relative is obtained by using arithmetic mean and geometric mean. When arithmetic mean is used for average of price relative, the formula for computing the index is

Simple average of price relative by arithmetic mean

$$P_{01} = \frac{\Sigma\left(\frac{p_1}{p_0} \times 100\right)}{n}$$

P_1 = Prices of current year
P_0 = Prices of base year
n = Number of items or commodities
when geometric mean is used for average of price relative, the formula for obtaining the index is
Simple average of price relative by geometric Mean

$$P_{01} = \text{Antilog}\left(\frac{\Sigma \log\left(\frac{p_1}{p_0} \times 100\right)}{n}\right)$$

Example 2:
From the following data, construct an index for 1998 taking 1997 as base by the average of price relative using (a) arithmetic mean and (b) Geometric mean

Commodity	Price in 1997	Price in 1998
A	50	70
B	40	60
C	80	100
D	20	30

Solution:
(a) Price relative index number using arithmetic mean

Commodity	Price in 1997 (P_0)	Price in 1998 (P_1)	$\frac{p_1}{p_0} \times 100$
A	50	70	140
B	40	60	150
C	80	100	125
D	20	30	150
		Total	565

Simple average of price relative index = (P_{01}) = $\frac{\Sigma\left(\frac{p_1}{p_0}\times 100\right)}{4}$

$= \frac{565}{4} = 141.25$

(b) Price relative index number using Geometric Mean

Commodity	Price in 1997 (P_0)	Price in 1998 (P_1)	$\frac{p_1}{p_0}\times 100$	$\log(\frac{p_1}{p_0}\times 100)$
A	50	70	140	2.1461
B	40	60	150	2.1761
C	80	100	125	2.0969
D	20	30	150	2.1761
			Total	8.5952

Simple average of price Relative index

(P_{01}) = Antilog $\frac{\Sigma \log\left[\frac{p_1}{p_o} x100\right]}{n}$

= Antilog $\frac{8.5952}{4}$

= Antilog [2.1488] = 140.9

Weighted aggregate index numbers

In order to attribute appropriate importance to each of the items used in an aggregate index number some reasonable weights must be used. There are various methods of assigning weights and consequently a large number of formulae for constructing index numbers have been devised of which some of the most important ones are

1. **Laspeyre' s method**
2. **Paasche' s method**
3. **Fisher' s ideal Method**
4. **Bowley' s Method**
5. **Marshall- Edgeworth method**
6. **Kelly' s Method**

1. Laspeyre' s method:

The Laspeyres price index is a weighted aggregate price index, where the weights are determined by quantities in the based period and is given by

$$\text{Laspeyre' s price index} = P_{01}^{L} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

2. Paasche' s method

The Paasche' s price index is a weighted aggregate price index in which the weight are determined by the quantities in the current year. The formulae for constructing the index is

$$\text{Paasche' s price index number} = P_{01}^{P} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

Where

P_0 = Price for the base year P_1 = Price for the current year

q_0 = Quantity for the base year q_1 = Quantity for the current year

3. Fisher' s ideal Method

Fisher' s Price index number is the geometric mean of the Laspeyres and Paasche indices Symbolically

$$\text{Fisher' s ideal index number} = P_{01}^{F} = \sqrt{L \times P}$$

$$= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

It is known as ideal index number because

(a) It is based on the geometric mean
(b) It is based on the current year as well as the base year
(c) It conform certain tests of consistency
(d) It is free from bias.

4. Bowley' s Method:

Bowley' s price index number is the arithmetic mean of Laspeyre' s and Paasche' s method. Symbolically

$$\text{Bowley' s price index number} = P_{01}^{B} = \frac{L + P}{2}$$

$$= \frac{1}{2}\left[\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right] \times 100$$

5. Marshall- Edgeworth method

This method also both the current year as well as base year prices and quantities are considered. The formula for constructing the index is

$$\text{Marshall Edgeworth price index} = P_{01}^{ME} = \frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} \times 100$$

$$= \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

6. Kelly' s Method

Kelly has suggested the following formula for constructing the index number

$$\text{Kelly' s Price index number} = P_{01}^{k} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

$$\text{Where} = q = \frac{q_0 + q_1}{2}$$

Here the average of the quantities of two years is used as weights

Example 3:

Construct price index number from the following data by applying

1. Laspeyere' s Method
2. Paasche' s Method
3. Fisher' s ideal Method

Commodity	2000		2001	
	Price	Qty	Price	Qty
A	2	8	4	5
B	5	12	6	10
C	4	15	5	12
D	2	18	4	20

Solution:

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	2	8	4	5	16	10	32	20
B	5	12	6	10	60	50	72	60
C	4	15	5	12	60	48	75	60
D	2	18	4	20	36	40	72	80
					172	148	251	220

Laspeyre' s price index = $P_{01}^{L} = \dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

$= \dfrac{251}{172} \times 100 \quad = 145.93$

Paasche price index number = $P_{01}^{P} = \dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$

$= \dfrac{220}{148} \times 100$

$= 148.7$

Fisher' s ideal index number $= \sqrt{L \times P}$

$= \sqrt{(145.9) \times (148.7)}$

$= \sqrt{21695.33}$

$= 147.3$

Or

Fisher' s ideal index number $= \sqrt{\dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$

$= \sqrt{\dfrac{251}{172} \times \dfrac{220}{148}} \times 100$

$= \sqrt{(1.459) \times (1.487)} \times 100$

$= \sqrt{2.170} \times 100$

$= 1.473 \times 100 = 147.3$

Interpretation:

The results can be interpreted as follows:

If 100 rupees were used in the base year to buy the given commodities, we have to use Rs 145.90 in the current year to buy the same amount of the commodities as per the Laspeyre' s formula. Other values give similar meaning .

Example 4:

Calculate the index number from the following data by applying

(a) Bowley' s price index

(b) Marshall- Edgeworth price index

Commodity	Base year		Current year	
	Quantity	Price	Quantity	Price
A	10	3	8	4
B	20	15	15	20
C	2	25	3	30

Solution:

Commodity	q_0	P_0	q_1	P_1	p_0q_0	p_0q_1	p_1q_0	$p_1 q_1$
A	10	3	8	4	30	24	40	32
B	20	15	15	20	300	225	400	300
C	2	25	3	30	50	75	60	90
					380	324	500	422

(a) Bowley' s price index number $= \frac{1}{2}\left[\frac{\Sigma p_1q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1q_1}{\Sigma p_0 q_1}\right] \times 100$

$= \frac{1}{2}\left[\frac{500}{380} + \frac{422}{324}\right] \times 100$

$= \frac{1}{2}\left[1.316 + 1.302\right] \times 100$

$= \frac{1}{2}\left[2.168\right] \times 100$

$= 1.309 \times 100$

$= 130.9$

(b) Marshall Edgeworths price index Number

$= P_{01}^{ME} = \frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} \times 100$

$= \left[\frac{500}{380} + \frac{422}{324}\right] \times 100$

$= \left[\frac{922}{704}\right] \times 100$

= 131. 0

Example 5:

Calculate a suitable price index from the following data

Commodity	Quantity	Price	
		1996	1997
A	20	2	4
B	15	5	6
C	8	3	2

Solution:

Here the quantities are given in common we can use Kelly' s index price number and is given by

$$\text{Kelly' s Price index number} = P_{01}{}^{k} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

$$= \frac{186}{139} \times 100 = 133.81$$

Commodity	q	P_0	P_1	p_0q	P_1q
A	20	2	4	40	80
B	15	5	6	75	90
C	8	3	2	24	16
			Total	139	186

$$\text{Kelly' s Price index number} = P_{01}{}^{k} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

IV. Weighted Average of Price Relative index.

When the specific weights are given for each commodity, the weighted index number is calculated by the formula.

$$\text{Weighted Average of Price Relative index} = \frac{\Sigma pw}{\Sigma w}$$

Where w = the weight of the commodity

P = the price relative index

$$= \frac{p_1}{p_0} \times 100$$

When the base year value P_0q_0 is taken as the weight i.e. $W=P_0q_0$ then the formula is

$$\text{Weighted Average of Price Relative index} = \frac{\Sigma\left(\frac{p_1}{p_0} \times 100\right) \times p_0q_0}{\Sigma p_0q_0}$$

$$= \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$$

This is nothing but Laspeyre' s formula.

When the weights are taken as $w = p_0q_1$, the formula is

$$\text{Weighted Average of Price Relative index} = \frac{\Sigma\left(\frac{p_1}{p_0} \times 100\right) \times p_0q_1}{\Sigma p_0q_1}$$

$$= \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100$$

This is nothing but Paasche' s Formula.

Example 6:

Compute the weighted index number for the following data.

Commodity	Price		Weight
	Current year	Base year	
A	5	4	60
B	3	2	50
C	2	1	30

Solution:

Commodity	P_1	P_0	W	$P=\frac{p_1}{p_0} \times 100$	PW
A	5	4	60	125	7500
B	3	2	50	150	7500

C	2	1	30	200	6000
			140		21000

Weighted Average of Price Relative index = $\frac{\Sigma pw}{\Sigma w}$

$= \frac{21000}{140}$

$= 150$

10.6 Quantity or Volume index number:

Price index numbers measure and permit comparison of the price of certain goods. On the other hand, the quantity index numbers measure the physical volume of production, employment and etc. The most common type of the quantity index is that of quantity produced.

Laspeyre' s quantity index number = $Q_{01}{}^{L} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$

Paasche' s quantity index number = $Q_{01}{}^{P} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100$

Fisher' s quantity index number = $Q_{01}{}^{F} = \sqrt{L \times P}$

$= \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100$

These formulae represent the quantity index in which quantities of the different commodities are weighted by their prices.

Example 7:

From the following data compute quantity indices by
(i) Laspeyre' s method, (ii) Paasche' s method and (iii) Fisher' s method.

	2000		2002	
Commodity	Price	Total value	Price	Total value
A	10	100	12	180
B	12	240	15	450

C	15	225	17	340

Solution:

Here instead of quantity, total values are given. Hence first find quantities of base year and current year,

ie. Quantity = $\frac{\text{total value}}{\text{price}}$

Commodity	p_0	q_0	P_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	10	10	12	15	100	150	120	180
B	12	20	15	30	240	360	300	450
C	15	15	17	20	225	300	255	340
					565	810	675	970

Laspeyre' s quantity index number = $q_{01}{}^L = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$

$= \frac{810}{565} \times 100$

$= 143.4$

Paasche' s quantity index number = $q_{01}{}^P = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100$

$= \frac{970}{675} \times 100$

$= 143.7$

Fisher' s quantity index number = $q_{01}{}^F = \sqrt{L \times P}$

$= \sqrt{143.4 \times 143.7}$

$= 143.6$

(or)

$$q_{01}{}^F = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100$$

$$= \sqrt{\frac{810}{565} \times \frac{970}{675}} \times 100$$

$$= \sqrt{1.434 \times 1.437} \times 100$$

$= 1.436 \times 100$

$= 143.6$

10.7 Tests of Consistency of index numbers:

Several formulae have been studied for the construction of index number. The question arises as to which formula is appropriate to a given problems. A number of tests been developed and the important among these are

1. Unit test
2. Time Reversal test
3. Factor Reversal test

1. Unit test:

The unit test requires that the formula for constructing an index should be independent of the units in which prices and quantities are quoted. Except for the simple aggregate index (unweighted) , all other formulae discussed in this chapter satisfy this test.

2. Time Reversal test:

Time Reversal test is a test to determine whether a given method will work both ways in time, forward and backward. In the words of Fisher, "the formula for calculating the index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as base". Symbolically, the following relation should be satisfied.

$P_{01} \times P_{10} = 1$

Where P_{01} is the index for time ' 1' as time ' 0' as base and P_{10} is the index for time ' 0' as time ' 1' as base. If the product is not unity, there is said to be a time bias is the method. Fisher' s ideal index satisfies the time reversal test.

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

Then $P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$

$= 1\sqrt{=} 1$

Therefore Fisher ideal index satisfies the time reversal test.

3. Factor Reversal test:

Another test suggested by Fisher is known s factor reversal test. It holds that the product of a price index and the quantity index should be equal to the corresponding value index. In the words of Fisher, "Just as each formula should permit the interchange of the two times without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent result, ie, the two results multiplied together should give the true value ratio.

In other word, if P_{01} represent the changes in price in the current year and Q_{01} represent the changes in quantity in the current year, then

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Thus based on this test, if the product is not equal to the value ratio, there is an error in one or both of the index number. The Factor reversal test is satisfied by the Fisher' s ideal index.

ie. $$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

Then $$P_{01} \times Q_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$= \sqrt{\left(\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}\right)^2}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Since $P_{01} \times Q_{01} = \dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$, the factor reversal test is satisfied by the Fisher' s ideal index.

Example 8:

Construct Fisher' s ideal index for the Following data. Test whether it satisfies time reversal test and factor reversal test.

	Base year		Current year	
Commodity	Quantity	Price	Quantity	Price
A	12	10	15	12
B	15	7	20	5
C	5	5	8	9

Solution:

Commodity	q_0	p_0	q_1	p_1	P_0q_0	p_0q_1	p_1q_0	p_1q_1
A	12	10	15	12	120	150	144	180
B	15	7	20	5	105	140	75	100
C	5	5	8	9	25	40	45	72
					250	330	264	352

Fisher ideal index number $P_{01}^F = \sqrt{\dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$

$$= \sqrt{\frac{264}{250} \times \frac{352}{330}} \times 100$$

$$= \sqrt{(1.056) \times (1.067)} \times 100$$

$$= \sqrt{1.127} \times 100$$

$$= 1.062 \times 100 = 106.2$$

Time Reversal test:

Time Reversal test is satisfied when $P_{01} \times P_{10} = 1$

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

$$= \sqrt{\frac{264}{250} \times \frac{352}{330}}$$

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$= \sqrt{\frac{330}{352} \times \frac{250}{264}}$$

Now $P_{01} \times P_{10} = \sqrt{\frac{264}{250} \times \frac{352}{330} \times \frac{330}{352} \times \frac{250}{264}}$

$= \sqrt{1}$

$= 1$

Hence Fisher ideal index satisfy the time reversal test.

Factor Reversal test:

Factor Reversal test is satisfied when $P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$

Now $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

$= \sqrt{\frac{264}{250} \times \frac{352}{330}}$

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$= \sqrt{\frac{330}{250} \times \frac{352}{264}}$$

Then $P_{01} \times Q_{01} = \sqrt{\frac{264}{250} \times \frac{352}{330} \times \frac{330}{250} \times \frac{352}{264}}$

$$= \sqrt{\left(\frac{352}{250}\right)^2}$$

$$= \frac{352}{250}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Hence Fisher ideal index number satisfy the factor reversal test.

10.8 Consumer Price Index

Consumer Price index is also called the cost of living index. It represent the average change over time in the prices paid by the ultimate consumer of a specified basket of goods and services. A change in the price level affects the costs of living of different classes of people differently. The general index number fails to reveal this. So there is the need to construct consumer price index. People consume different types of commodities. People' s consumption habit is also different from man to man, place to place and class to class i.e richer class, middle class and poor class.

The scope of consumer price is necessary, to specify the population group covered. For example, working class, poor class, middle class, richer class, etc and the geographical areas must be covered as urban, rural, town, city etc.

Use of Consumer Price index

The consumer price indices are of great significance and is given below.

1. This is very useful in wage negotiations, wage contracts and dearness allowance adjustment in many countries.
2. At government level, the index numbers are used for wage policy, price policy, rent control, taxation and general economic policies.
3. Change in the purchasing power of money and real income can be measured.
4. Index numbers are also used for analysing market price for particular kinds of goods and services.

Method of Constructing Consumer price index:

There are two methods of constructing consumer price index. They are

1. Aggregate Expenditure method (or) Aggregate method.
2. Family Budget method (or) Method of Weighted Relative method.

1. Aggregate Expenditure method:

This method is based upon the Laspeyre' s method. It is widely used. The quantities of commodities consumed by a particular group in the base year are the weight.

The formula is Consumer Price Index number = $\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

2. Family Budget method or Method of Weighted Relatives:

This method is estimated an aggregate expenditure of an average family on various items and it is weighted. The formula is

Consumer Price index number = $\frac{\Sigma pw}{\Sigma w}$

Where P = $\frac{p_1}{p_0} \times 100$ for each item. w = value weight (i.e) $p_0 q_0$

"Weighted average price relative method" which we have studied before and "Family Budget method" are the same for finding out consumer price index.

Example 9:

Construct the consumer price index number for 1996 on the basis of 1993 from the following data using Aggregate expenditure method.

		Price in	
Commodity	Quantity consumed	1993	1996
A	100	8	12
B	25	6	7
C	10	5	8
D	20	15	18

Solution:

Commodity	q_0	p_0	p_1	p_0q_0	p_1q_0
A	100	8	12	800	1200
B	25	6	7	150	175
C	10	5	8	50	80
D	20	15	18	300	360
			Total	1300	1815

Consumer price index by Aggregate expenditure method

$$= \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

$$= \frac{1815}{1300} \times 100 = 139.6$$

Example 10:

Calculate consumer price index by using Family Budget method for year 1993 with 1990 as base year from the following data.

		Price in	
Items	Weights	1990 (Rs.)	1993 (Rs.)
Food	35	150	140
Rent	20	75	90
Clothing	10	25	30
Fuel and lighting	15	50	60
Miscellaneous	20	60	80

Solution:

Items	W	P_0	P_1	$P = \frac{p_1}{p_0} \times 100$	PW
Food	35	150	140	93.33	3266.55
Rent	20	75	90	120.00	2400.00
Clothing	10	25	30	120.00	1200.00
Fuel and lighting	15	50	60	120.00	1800.00
Miscellaneous	20	60	80	133.33	2666.60
	100				11333.15

$$\text{Consumer price index by Family Budget method} = \frac{\Sigma pw}{\Sigma w}$$

$$= \frac{11333.15}{100}$$

$$= 113.33$$

Exercise – 10

I. Choose the correct answer:

1. Index number is a
   (a) measure of relative changes
   (b) a special type of an average
   (c) a percentage relative
   (d) all the above
2. Most preferred type of average for index number is
   (a) arithmetic mean
   (b) geometric mean
   (c) hormonic mean
   (d) none of the above
3. Laspeyre' s index formula uses the weights of the
   (a) base year
   (b) current year
   (c) average of the weights of a number of years
   (d) none of the above
4. The geometric mean of Laspeyere' s and Passche' s price indices is also known as
   (a) Fisher' s price index
   (b) Kelly' s price index
   (c) Marshal-Edgeworth index number
   (d) Bowley' s price index
5. The condition for the time reversal test to hold good with usual notations is
   (a) $P_{01} \times P_{10} = 1$
   (b) $P_{10} \times P_{01} = 0$
   (c) $P_{01} / P_{10} = 1$
   (d) $P_{01} + P_{10} = 1$
6. An appropriate method for working out consumer price index is
   (a) weighted aggregate expenditure method
   (b) family budget method
   (c) price relative method
   (d) none of the above

7. The weights used in Passche' s formula belong to
   (a) The base period
   (b) The given period
   (c) To any arbitrary chosen period
   (d) None of the above

II. Fill in the blank in the following

8. Index numbers help in framing of _____
9. Fisher' s ideal index number is the_____of Laspeyer's and Paasche' s index numbers
10. Index numbers are expressed in _____
11. _____is known as Ideal index number
12. In family budget method, the cost of living index number is _____

III. Answer the following

13. What is an index number? What are the uses of index numbers.
14. Explain Time Reversal Test and Factor Reversal test.
15. What is meant by consumer price index number? What are its uses.
16. Calculate price index number by
    (i) Laspeyre' s method
    (ii) Paasche' s method
    (iii) Fisher' s ideal index method.

	1990		1995	
Commodity	Price	Quantity	Price	Quantity
A	20	15	30	20
B	15	10	20	15
C	30	20	25	10
D	10	5	12	10

17. Calculate Fisher ideal index for the following data. Also test whether it satisfies time reversal test and factor reversal test.

	Price		Quantity	
Commodity	2000	2002	2000	2002
A	6	35	10	40

B	10	25	12	30
C	12	15	8	20

18. Calculate the cost of living index number from the following data.

	Price		
Items	Base year	Current year	Weight
Food	30	45	4
Fuel	10	15	2
Clothing	15	20	1
House Rent	20	15	3
Miscellaneous	25	20	2

Answers

I.

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

II.

8. Polices 9. Geometric mean 10. Percentage
11. Fisher' s index number 12. $\frac{\Sigma pw}{\Sigma w}$

III.

16. (i) L = 110
    (ii) P = 123.9
    (iii) F = 116.7
17. 296
18. 118.2