OBJECTIVE: The aim of the present lesson is to enable the students to understand the meaning, definition, nature, importance and limitations of statistics.

> *"A knowledge of statistics is like a knowledge of foreign language of algebra; it may prove of use at any time under any circumstance"……………………………………......Bowley.*

STRUCTURE:

1.1 INTRODUCTION

For a layman, 'Statistics' means numerical information expressed in quantitative terms. This information may relate to objects, subjects, activities, phenomena, or regions of space. As a matter of fact, data have no limits as to their reference, coverage, and scope. At the macro level, these are data on gross national product and shares of agriculture, manufacturing, and services in GDP (Gross Domestic Product).

At the micro level, individual firms, howsoever small or large, produce extensive statistics on their operations. The annual reports of companies contain variety of data on sales, production, expenditure, inventories, capital employed, and other activities. These data are often field data, collected by employing scientific survey techniques. Unless regularly updated, such data are the product of a one-time effort and have limited use beyond the situation that may have called for their collection. A student knows statistics more intimately as a subject of study like economics, mathematics, chemistry, physics, and others. It is a discipline, which scientifically deals with data, and is often described as the science of data. In dealing with statistics as data, statistics has developed appropriate methods of collecting, presenting, summarizing, and analysing data, and thus consists of a body of these methods.

1.2 MEANING AND DEFINITIONS OF STATISTICS

In the beginning, it may be noted that the word 'statistics' is used rather curiously in two senses plural and singular. In the plural sense, it refers to a set of figures or data. In the singular sense, statistics refers to the whole body of tools that are used to collect data, organise and interpret them and, finally, to draw conclusions from them. It should be noted that both the aspects of statistics are important if the quantitative data are to serve their purpose. If statistics, as a subject, is inadequate and consists of poor methodology, we could not know the right procedure to extract from the data the information they contain. Similarly, if our data are defective or that they are inadequate or inaccurate, we could not reach the right conclusions even though our subject is well developed.

A.L. Bowley has defined statistics as: (i) statistics is the science of counting, (ii) Statistics may rightly be called the science of averages, and (iii) statistics is the science of measurement of social organism regarded as a whole in all its manifestations. *Boddington* defined as: Statistics is the science of estimates and probabilities. Further, *W.I. King* has defined Statistics in a wider context, the science of Statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis or enumeration or collection of estimates.

Seligman explored that statistics is a science that deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry. *Spiegal* defines statistics highlighting its role in decision-making particularly under uncertainty, as follows: statistics is concerned with scientific method for collecting, organising, summa rising, presenting and analyzing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis. According to *Prof. Horace Secrist*, Statistics is the aggregate of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose, and placed in relation to each other.

From the above definitions, we can highlight the major characteristics of statistics as follows:

(i) *Statistics are the aggregates of facts*. It means a single figure is not statistics. For example, national income of a country for a single year is not statistics but the same for two or more years is statistics.

(ii) *Statistics are affected by a number of factors.* For example, sale of a product depends on a number of factors such as its price, quality, competition, the income of the consumers, and so on.

(iii) *Statistics must be reasonably accurate.* Wrong figures, if analysed, will lead to erroneous conclusions. Hence, it is necessary that conclusions must be based on accurate figures.

(iv) *Statistics must be collected in a systematic manner.* If data are collected in a haphazard manner, they will not be reliable and will lead to misleading conclusions.

(v) *Collected in a systematic manner for a pre-determined purpose*

(vi) Lastly, Statistics should be placed in relation to each other. If one collects data unrelated to each other, then such data will be confusing and will not lead to any logical conclusions. Data should be comparable over time and over space.

1.3 TYPES OF DATA AND DATA SOURCES

Statistical data are the basic raw material of statistics. Data may relate to an activity of our interest, a phenomenon, or a problem situation under study. They derive as a result of the process of measuring, counting and/or observing. Statistical data, therefore, refer to those aspects of a problem situation that can be measured, quantified, counted, or classified. Any object subject phenomenon, or activity that generates data through this process is termed as a variable. In other words, a variable is one that shows a degree of variability when successive measurements are recorded. In statistics, data are classified into two broad categories: quantitative data and qualitative data. This classification is based on the kind of characteristics that are measured.

Quantitative data are those that can be quantified in definite units of measurement. These refer to characteristics whose successive measurements yield quantifiable observations. Depending on the nature of the variable observed for measurement, quantitative data can be further categorized as continuous and discrete data.

Obviously, a variable may be a continuous variable or a discrete variable.

(i) **Continuous data** represent the numerical values of a continuous variable. A continuous variable is the one that can assume any value between any two points on a line segment, thus representing an interval of values. The values are quite precise and close to each other, yet distinguishably different. All characteristics such as weight, length, height, thickness, velocity, temperature, tensile strength, etc., represent continuous variables. Thus, the data recorded on these and similar other characteristics are called continuous data. It may be noted that a continuous variable assumes the finest unit of measurement. Finest in the sense that it enables measurements to the maximum degree of precision.

(ii) **Discrete data** are the values assumed by a discrete variable. A discrete variable is the one whose outcomes are measured in fixed numbers. Such data are essentially count data. These are derived from a process of counting, such as the number of items possessing or not possessing a certain characteristic. The number of customers visiting a departmental store everyday, the incoming flights at an airport, and the defective items in a consignment received for sale, are all examples of discrete data.

Qualitative data refer to qualitative characteristics of a subject or an object. A characteristic is qualitative in nature when its observations are defined and noted in terms of the presence or absence of a certain attribute in discrete numbers. These data are further classified as nominal and rank data.

(i) **Nominal data** are the outcome of classification into two or more categories of items or units comprising a sample or a population according to some quality characteristic. Classification of students according to sex (as males and females), of workers according to skill (as skilled, semi-skilled, and unskilled), and of employees according to the level of education (as matriculates, undergraduates, and post-graduates), all result into nominal data. Given any such basis of classification, it is always possible to assign each item to a particular class and make a summation of items belonging to each class. The count data so obtained are called nominal data.

(ii) Rank data, on the other hand, are the result of assigning ranks to specify order in terms of the integers 1,2,3, ..., n. Ranks may be assigned according to the level of performance in a test. a contest, a competition, an interview, or a show. The candidates appearing in an interview, for example, may be assigned ranks in integers ranging from I to n, depending on their performance in the interview. Ranks so assigned can be viewed as the continuous values of a variable involving performance as the quality characteristic.

Data sources could be seen as of two types, viz., secondary and primary. The two can be defined as under:

(i) **Secondary data:** They already exist in some form: published or unpublished - in an identifiable secondary source. They are, generally, available from published source(s), though not necessarily in the form actually required.

(ii) **Primary data:** Those data which do not already exist in any form, and thus have to be collected for the first time from the primary source(s). By their very nature, these data require fresh and first-time collection covering the whole population or a sample drawn from it.

1.4 TYPES OF STATISTICS

There are two major divisions of statistics such as descriptive statistics and inferential statistics. The term **descriptive statistics** deals with collecting, summarizing, and simplifying data, which are otherwise quite unwieldy and voluminous. It seeks to achieve this in a manner that meaningful conclusions can be readily drawn from the data. Descriptive statistics may thus be seen as comprising methods of bringing out and highlighting the latent characteristics present in a set of numerical data. It not only facilitates an understanding of the data and systematic reporting thereof in a manner; and also makes them amenable to further discussion, analysis, and interpretations.

The first step in any scientific inquiry is to collect data relevant to the problem in hand. When the inquiry relates to physical and/or biological sciences, data collection is normally an integral part of the experiment itself. In fact, the very manner in which an experiment is designed, determines the kind of data it would require and/or generate. The problem of identifying the nature and the kind of the relevant data is thus automatically resolved as soon as the design of experiment is finalized. It is possible in the case of physical sciences. In the case of social sciences, where the required data are often collected through a questionnaire from a number of carefully selected respondents, the problem is not that simply resolved. For one thing, designing the questionnaire itself is a critical initial problem. For another, the number of respondents to be accessed for data collection and the criteria for selecting them has their own implications and importance for the quality of results obtained. Further, the data have been collected, these are assembled, organized, and presented in the form of appropriate tables to make them readable. Wherever needed, figures, diagrams, charts, and graphs are also used for better presentation of the data. A useful tabular and graphic presentation of data will require that the raw data be properly classified in accordance with the objectives of investigation and the relational analysis to be carried out. .

A well thought-out and sharp data classification facilitates easy description of the hidden data characteristics by means of a variety of summary measures. These include measures of central tendency, dispersion, skewness, and kurtosis, which constitute the essential scope of descriptive statistics. These form a large part of the subject matter of any basic textbook on the subject, and thus they are being discussed in that order here as well.

Inferential statistics, also known as inductive statistics, goes beyond describing a given problem situation by means of collecting, summarizing, and meaningfully presenting the related data. Instead, it consists of methods that are used for drawing inferences, or making broad generalizations, about a totality of observations on the basis of knowledge about a part of that totality. The totality of observations about which an inference may be drawn, or a generalization made, is called a population or a universe. The part of totality, which is observed for data collection and analysis to gain knowledge about the population, is called a sample.

The desired information about a given population of our interest; may also be collected even by observing all the units comprising the population. This total coverage is called census. Getting the desired value for the population through census is not always feasible and practical for various reasons. Apart from time and money considerations making the census operations prohibitive, observing each individual unit of the population with reference to any data characteristic may at times involve even destructive testing. In such cases, obviously, the only recourse available is to employ the partial or incomplete information gathered through a sample for the purpose. This is precisely what inferential statistics does. Thus, obtaining a particular value from the sample information and using it for drawing an inference about the entire population underlies the subject matter of inferential statistics. Consider a situation in which one is required to know the average body weight of all the college students in a given cosmopolitan city during a certain year. A quick and easy way to do this is to record the weight of only 500 students, from out of a total strength of, say, 10000, or an unknown total strength, take the average, and use this average based on incomplete weight data to represent the average body weight of all the college students. In a different situation, one may have to repeat this exercise for some future year and use the quick estimate of average body weight for a comparison. This may be needed, for example, to decide whether the weight of the college students has undergone a significant change over the years compared.

Inferential statistics helps to evaluate the risks involved in reaching inferences or generalizations about an unknown population on the basis of sample information. for example, an inspection of a sample of five battery cells drawn from a given lot may reveal that all the five cells are in perfectly good condition. This information may be used to conclude that the entire lot is good enough to buy or not.

Since this inference is based on the examination of a sample of limited number of cells, it is equally likely that all the cells in the lot are not in order. It is also possible that all the items that may be included in the sample are unsatisfactory. This may be used to conclude that the entire lot is of unsatisfactory quality, whereas the fact may indeed be otherwise. It may, thus, be noticed that there is always a risk of an inference about a population being incorrect when based on the knowledge of a limited sample. The rescue in such situations lies in evaluating such risks. For this, statistics provides the necessary methods. These centres on quantifying in probabilistic term the chances of decisions taken on the basis of sample information being incorrect. This requires an understanding of the what, why, and how of probability and probability distributions to equip ourselves with methods of drawing statistical inferences and estimating the degree of reliability of these inferences.

1.5 SCOPE OF STATISTICS

Apart from the methods comprising the scope of descriptive and inferential branches of statistics, statistics also consists of methods of dealing with a few other issues of specific nature. Since these methods are essentially descriptive in nature, they have been discussed here as part of the descriptive statistics. These are mainly concerned with the following:

(i) It often becomes necessary to examine how two paired data sets are related. For example, we may have data on the sales of a product and the expenditure incurred on its advertisement for a specified number of years. Given that sales and advertisement expenditure are related to each other, it is useful to examine the nature of relationship between the two and quantify the degree of that relationship. As this requires use of appropriate statistical methods, these falls under the purview of what we call regression and correlation analysis.

(ii) Situations occur quite often when we require averaging (or totalling) of data on prices and/or quantities expressed in different units of measurement. For example, price of cloth may be quoted per meter of length and that of wheat per kilogram of weight. Since ordinary methods of totalling and averaging do not apply to such price/quantity data, special techniques needed for the purpose are developed under index numbers.

(iii) Many a time, it becomes necessary to examine the past performance of an activity with a view to determining its future behaviour. For example, when engaged in the production of a commodity, monthly product sales are an important measure of evaluating performance. This requires compilation and analysis of relevant sales data over time. The more complex the activity, the more varied the data requirements. For profit maximising and future sales planning, forecast of likely sales growth rate is crucial. This needs careful collection and analysis of past sales data. All such concerns are taken care of under time series analysis.

(iv) Obtaining the most likely future estimates on any aspect(s) relating to a business or economic activity has indeed been engaging the minds of all concerned. This is particularly important when it relates to product sales and demand, which serve the necessary basis of production scheduling and planning. The regression, correlation, and time series analyses together help develop the basic methodology to do the needful. Thus, the study of methods and techniques of obtaining the likely estimates on business/economic variables comprises the scope of what we do under business forecasting.

Keeping in view the importance of inferential statistics, the scope of statistics may finally be restated as consisting of statistical methods which facilitate decision--making under conditions of uncertainty. While the term statistical methods is often used to cover the subject of statistics as a whole, in particular it refers to methods by which statistical data are analysed, interpreted, and the inferences drawn for decision-making.

Though generic in nature and versatile in their applications, statistical methods have come to be widely used, especially in all matters concerning business and economics. These are also being increasingly used in biology, medicine, agriculture, psychology, and education. The scope of application of these methods has started opening and expanding in a number of social science disciplines as well. Even a political scientist finds them of increasing relevance for examining the political behaviour and it is, of course, no surprise to find even historians statistical data, for history is essentially past data presented in certain actual format.

1.6 IMPORTANCE OF STATISTICS IN BUSINESS

There are three major functions in any business enterprise in which the statistical methods are useful. These are as follows:

(i) **The planning of operations:** This may relate to either special projects or to the recurring activities of a firm over a specified period.

(ii) **The setting up of standards:** This may relate to the size of employment, volume of sales, fixation of quality norms for the manufactured product, norms for the daily output, and so forth.

(iii) **The function of control:** This involves comparison of actual production achieved against the norm or target set earlier. In case the production has fallen short of the target, it gives remedial measures so that such a deficiency does not occur again.

A worth noting point is that although these three functions-planning of operations, setting standards, and control-are separate, but in practice they are very much interrelated.

Different authors have highlighted the importance of Statistics in business. For instance, Croxton and Cowden give numerous uses of Statistics in business such as project planning, budgetary planning and control, inventory planning and control, quality control, marketing, production and personnel administration. Within these also they have specified certain areas where Statistics is very relevant. Another author, Irwing W. Burr, dealing with the place of statistics in an industrial organisation, specifies a number of areas where statistics is extremely useful. These are: customer wants and market research, development design and specification, purchasing, production, inspection, packaging and shipping, sales and complaints, inventory and maintenance, costs, management control, industrial engineering and research.

Statistical problems arising in the course of business operations are multitudinous. As such, one may do no more than highlight some of the more important ones to emphasis the relevance of statistics to the business world. In the sphere of production, for example, statistics can be useful in various ways.

Statistical quality control methods are used to ensure the production of quality goods. Identifying and rejecting defective or substandard goods achieve this. The sale targets can be fixed on the basis of sale forecasts, which are done by using varying methods of forecasting. Analysis of sales affected against the targets set earlier would indicate the deficiency in achievement, which may be on account of several causes: (i) targets were too high and unrealistic (ii) salesmen's performance has been poor (iii) emergence of increase in competition (iv) poor quality of company's product, and so on. These factors can be further investigated.

Another sphere in business where statistical methods can be used is personnel management. Here, one is concerned with the fixation of wage rates, incentive norms and performance appraisal of individual employee. The concept of productivity is very relevant here. On the basis of measurement of productivity, the productivity bonus is awarded to the workers. Comparisons of wages and productivity are undertaken in order to ensure increases in industrial productivity.

Statistical methods could also be used to ascertain the efficacy of a certain product, say, medicine. For example, a pharmaceutical company has developed a new medicine in the treatment of bronchial asthma. Before launching it on commercial basis, it wants to ascertain the effectiveness of this medicine. It undertakes an experimentation involving the formation of two comparable groups of asthma patients. One group is given this new medicine for a specified period and the other one is treated with the usual medicines. Records are maintained for the two groups for the specified period. This record is then analysed to ascertain if there is any significant difference in the recovery of the two groups. If the difference is really significant statistically, the new medicine is commercially launched.

1.7 LIMITATIONS OF STATISTICS

Statistics has a number of limitations, pertinent among them are as follows:

(i) There are certain phenomena or concepts where statistics cannot be used. This is because these phenomena or concepts are not amenable to measurement. For example, beauty, intelligence, courage cannot be quantified. Statistics has no place in all such cases where quantification is not possible.

(ii) Statistics reveal the average behaviour, the normal or the general trend. An application of the 'average' concept if applied to an individual or a particular situation may lead to a wrong conclusion and sometimes may be disastrous. For example, one may be misguided when told that the average depth of a river from one bank to the other is four feet, when there may be some points in between where its depth is far more than four feet. On this understanding, one may enter those points having greater depth, which may be hazardous.

(iii) Since statistics are collected for a particular purpose, such data may not be relevant or useful in other situations or cases. For example, secondary data (i.e., data originally collected by someone else) may not be useful for the other person.

(iv) Statistics are not 100 per cent precise as is Mathematics or Accountancy. Those who use statistics should be aware of this limitation.

(v) In statistical surveys, sampling is generally used as it is not physically possible to cover all the units or elements comprising the universe. The results may not be appropriate as far as the universe is concerned. Moreover, different surveys based on the same size of sample but different sample units may yield different results.

(vi) At times, association or relationship between two or more variables is studied in statistics, but such a relationship does not indicate cause and effect' relationship. It simply shows the similarity or dissimilarity in the movement of the two variables. In such cases, it is the user who has to interpret the results carefully, pointing out the type of relationship obtained.

(vii) A major limitation of statistics is that it does not reveal all pertaining to a certain phenomenon. There is some background information that statistics does not cover. Similarly, there are some other aspects related to the problem on hand, which are also not covered. The user of Statistics has to be well informed and should interpret Statistics keeping in mind all other aspects having relevance on the given problem.

Apart from the limitations of statistics mentioned above, there are misuses of it. Many people, knowingly or unknowingly, use statistical data in wrong manner. Let us see what the main misuses of statistics are so that the same could be avoided when one has to use statistical data. The misuse of Statistics may take several forms some of which are explained below.

(i) **Sources of data not given:** At times, the source of data is not given. In the absence of the source, the reader does not know how far the data are reliable. Further, if he wants to refer to the original source, he is unable to do so.

(ii) **Defective data:** Another misuse is that sometimes one gives defective data. This may be done knowingly in order to defend one's position or to prove a particular point. This apart, the definition used to denote a certain phenomenon may be defective. For example, in case of data relating to unemployed persons, the definition may include even those who are employed, though partially. The question here is how far it is justified to include partially employed persons amongst unemployed ones.

(iii) **Unrepresentative sample:** In statistics, several times one has to conduct a survey, which necessitates to choose a sample from the given population or universe. The sample may turn out to be unrepresentative of the universe. One may choose a sample just on the basis of convenience. He may collect the desired information from either his friends or nearby respondents in his neighbourhood even though such respondents do not constitute a representative sample.

(iv) **Inadequate sample:** Earlier, we have seen that a sample that is unrepresentative of the universe is a major misuse of statistics. This apart, at times one may conduct a survey based on an extremely inadequate sample. For example, in a city we may find that there are 1, 00,000 households. When we have to conduct a household survey, we may take a sample of merely 100 households comprising only 0.1 per cent of the universe. A survey based on such a small sample may not yield right information.

(v) **Unfair Comparisons:** An important misuse of statistics is making unfair comparisons from the data collected. For instance, one may construct an index of production choosing the base year where the production was much less. Then he may compare the subsequent year's production from this low base.

Such a comparison will undoubtedly give a rosy picture of the production though in reality it is not so. Another source of unfair comparisons could be when one makes absolute comparisons instead of relative ones. An absolute comparison of two figures, say, of production or export, may show a good increase, but in relative terms it may turnout to be very negligible. Another example of unfair comparison is when the population in two cities is different, but a comparison of overall death rates and deaths by a particular disease is attempted. Such a comparison is wrong. Likewise, when data are not properly classified or when changes in the composition of population in the two years are not taken into consideration, comparisons of such data would be unfair as they would lead to misleading conclusions.

(vi) **Unwanted conclusions:** Another misuse of statistics may be on account of unwarranted conclusions. This may be as a result of making false assumptions. For example, while making projections of population in the next five years, one may assume a lower rate of growth though the past two years indicate otherwise. Sometimes one may not be sure about the changes in business environment in the near future. In such a case, one may use an assumption that may turn out to be wrong. Another source of unwarranted conclusion may be the use of wrong average. Suppose in a series there are extreme values, one is too high while the other is too low, such as 800 and 50. The use of an arithmetic average in such a case may give a wrong idea. Instead, harmonic mean would be proper in such a case.

(vii) **Confusion of correlation and causation:** In statistics, several times one has to examine the relationship between two variables. A close relationship between the two variables may not establish a cause-and-effect-relationship in the sense that one variable is the cause and the other is the effect. It should be taken as something that measures degree of association rather than try to find out causal relationship..

OBJECTIVE: The present lesson imparts understanding of the calculations and main properties of measures of central tendency, including mean, mode, median, quartiles, percentiles, etc.

STRUCTURE:

2.1 INTRODUCTION

The description of statistical data may be quite elaborate or quite brief depending on two factors: the nature of data and the purpose for which the same data have been collected. While describing data statistically or verbally, one must ensure that the description is neither too brief nor too lengthy. The measures of central tendency enable us to compare two or more distributions pertaining to the same time period or within the same distribution over time. For example, the average consumption of tea in two different territories for the same period or in a territory for two years, say, 2003 and 2004, can be attempted by means of an average.

2.2 ARITHMETIC MEAN

Adding all the observations and dividing the sum by the number of observations results the arithmetic mean. Suppose we have the following observations:

10, 15,30, 7, 42, 79 and 83

These are seven observations. Symbolically, the arithmetic mean, also called simply *mean* is

$\bar{x}$ $= \sum x/n$, where $\bar{x}$ is simple mean.

$$= \frac{10+15+30+7+42+79+83}{7}$$

$$= \frac{266}{7} = 38$$

It may be noted that the Greek letter μ is used to denote the mean of the population and *n* to denote the total number of observations in a population. Thus the population mean $\mu = \sum x/n$. The formula given above is the basic formula that forms the definition of arithmetic mean and is used in case of ungrouped data where weights are not involved.

2.2.1 UNGROUPED DATA-WEIGHTED AVERAGE

In case of ungrouped data where weights are involved, our approach for calculating arithmetic mean will be different from the one used earlier.

Example 2.1: Suppose a student has secured the following marks in three tests:

Mid-term test	30
Laboratory	25
Final	20

The simple arithmetic mean will be $\frac{30+25+20}{3} = 25$

However, this will be wrong if the three tests carry different weights on the basis of their relative importance. Assuming that the weights assigned to the three tests are:

Mid-term test	2 points
Laboratory	3 points
Final	5 points

Solution: On the basis of this information, we can now calculate a weighted mean as shown below:

Table 2.1: Calculation of a Weighted Mean

Type of Test	Relative Weight (w)	Marks (x)	(wx)
Mid-term	2	30	60
Laboratory	3	25	75
Final	5	20	100
Total	∑ w = 10		235

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3}$$

$$= \frac{60 + 75 + 100}{2 + 3 + 5} = 23.5 \text{ marks}$$

It will be seen that weighted mean gives a more realistic picture than the simple or unweighted mean.

Example 2.2: An investor is fond of investing in equity shares. During a period of falling prices in the stock exchange, a stock is sold at Rs 120 per share on one day, Rs 105 on the next and Rs 90 on the third day. The investor has purchased 50 shares on the first day, 80 shares on the second day and 100 shares on the third' day. What average price per share did the investor pay?

Solution:

Table 2.2: Calculation of Weighted Average Price

Day	Price per Share (Rs) (x)	No of Shares Purchased (w)	Amount Paid (wx)
1	120	50	6000
2	105	80	8400
3	90	100	9000
Total	-	230	23,400

Weighted average $= \dfrac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3} = \dfrac{\sum wx}{\sum w}$

$= \dfrac{6000 + 8400 + 9000}{50 + 80 + 100} = 101.7 \text{ marks}$

Therefore, the investor paid an average price of Rs 101.7 per share.

It will be seen that if merely prices of the shares for the three days (regardless of the number of shares purchased) were taken into consideration, then the average price would be

$Rs. \dfrac{120 + 105 + 90}{3} = 105$

This is an unweighted or simple average and as it ignores the-quantum of shares purchased, it fails to give a correct picture. A simple average, it may be noted, is also a weighted average where weight in each case is the same, that is, only 1. When we use the term average alone, we always mean that it is an unweighted or simple average.

2.2.2 GROUPED DATA-ARITHMETIC MEAN

For grouped data, arithmetic mean may be calculated by applying any of the following methods:

(i) Direct method, (ii) Short-cut method ,(iii) Step-deviation method

In the case of direct method, the formula $\bar{x} = \sum fm/n$ is used. Here *m* is mid-point of various classes, *f* is the frequency of each class and *n* is the total number of frequencies. The calculation of arithmetic mean by the direct method is shown below.

Example 2.3: The following table gives the marks of 58 students in Statistics. Calculate the average marks of this group.

Marks	*No. of Students*
0-10	4
10-20	8
20-30	11
30-40	15
40-50	12
50-60	6
60-70	2
Total	58

Solution:

Table 2.3: Calculation of Arithmetic Mean by Direct Method

Marks	Mid-point m	No. of Students f	fm
0-10	5	4	20
10-20	15	8	120
20-30	25	11	275
30-40	35	15	525
40-50	45	12	540
50-60	55	6	330
60-70	65	2	130
			∑fm = 1940

Where,

$$\bar{x} = \frac{\sum fm}{n} = \frac{1940}{58} = \text{ 33.45 marks or 33 marks approximately.}$$

It may be noted that the mid-point of each class is taken as a good approximation of the true mean of the class. This is based on the assumption that the values are distributed fairly evenly throughout the interval. When large numbers of frequency occur, this assumption is usually accepted.

In the case of short-cut method, the concept of arbitrary mean is followed. The formula for calculation of the arithmetic mean by the short-cut method is given below:

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

Where A = arbitrary or assumed mean

f = frequency

d = deviation from the arbitrary or assumed mean

When the values are extremely large and/or in fractions, the use of the direct method would be very cumbersome. In such cases, the short-cut method is preferable. This is because the calculation work in the short-cut method is considerably reduced particularly for calculation of the product of values and their respective frequencies. However, when calculations are not made manually but by a machine calculator, it may not be necessary to resort to the short-cut method, as the use of the direct method may not pose any problem.

As can be seen from the formula used in the short-cut method, an arbitrary or assumed mean is used. The second term in the *formula* ($\sum fd \div$ n) is the correction factor for the difference between the actual mean and the assumed mean. If the assumed mean turns out to be equal to the actual mean, ($\sum fd \div$ n) will be zero. The use of the short-cut method is based on the principle that the total of deviations taken from an actual mean is equal to zero. As such, the deviations taken from any other figure will depend on how the assumed mean is related to the actual mean. While one may choose any value as assumed mean, it would be proper to avoid extreme values, that is, too small or too high to simplify calculations. A value apparently close to the arithmetic mean should be chosen.

For the figures given earlier pertaining to marks obtained by 58 students, we calculate the average marks by using the short-cut method.

Example 2.4:

Table 2.4: Calculation of Arithmetic Mean by Short-cut Method

Marks	Mid-point m	f	d	fd
0-10	5	4	-30	-120
10-20	15	8	-20	-160
20-30	25	11	-10	-110
30-40	35	15	0	0
40-50	45	12	10	120
50-60	55	6	20	120
60-70	65	2	30	60
				∑fd = -90

It may be noted that we have taken arbitrary mean as 35 and deviations from midpoints. In other words, the arbitrary mean has been subtracted from each value of mid-point and the resultant figure is shown in column *d.*

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

$$= 35 + \left(\frac{-90}{58}\right)$$

= 35 - 1.55 = 33.45 or 33 marks approximately.

Now we take up the calculation of arithmetic mean for the same set of data using the step-deviation method. This is shown in Table 2.5.

Table 2.5: Calculation of Arithmetic Mean by Step-deviation Method

Marks	Mid-point	f	d	d'= d/10	Fd'
0-10	5	4	-30	-3	-12
10-20	15	8	-20	-2	-16
20-30	25	11	-10	-1	-11
30-40	35	15	0	0	0
40-50	45	12	10	1	12
50-60	55	6	20	2	12
60-70	65	2	30	3	6
					∑fd' =-9

$$\bar{x} = A + \frac{\sum fd'}{n} \times C$$

$$= 35 + \left(\frac{-9 \times 10}{58} \right) \qquad = 33.45 \text{ or } 33 \text{ marks approximately.}$$

It will be seen that the answer in each of the three cases is the same. The step-deviation method is the most convenient on account of simplified calculations. It may also be noted that if we select a different arbitrary mean and recalculate deviations from that figure, we would get the same answer.

Now that we have learnt how the arithmetic mean can be calculated by using different methods, we are in a position to handle any problem where calculation of the arithmetic mean is involved.

Example 2.6: The mean of the following frequency distribution was found to be 1.46.

No. of Accidents	*No. of Days (frequency)*
0	46
1	?
2	?
3	25
4	10
5	5
Total	200 days

Calculate the missing frequencies.

Solution:

Here we are given the total number of frequencies and the arithmetic mean. We have to determine the two frequencies that are missing. Let us assume that the frequency against 1 accident is x and against 2 accidents is y. If we can establish two simultaneous equations, then we can easily find the values of X and *Y.*

$$\text{Mean} = \frac{(0.46) + (1.\text{x}) + (2.\text{y}) + (3.25) + (4.10) + (5.5)}{200}$$

$$1.46 = \frac{x + 2y + 140}{200}$$

$x + 2y + 140 = (200)\ (1.46)$

$x + 2y = 152$

$x + y = 200 - \{46+25 + 1O+5\}$

$x + y = 200 - 86$

$x + y = 114$

Now subtracting equation (ii) from equation (i), we get

$$\begin{array}{lcl} x + 2y & = & 152 \\ x + y & = & 114 \\ - \;\; - & & - \\ \hline y & = & 38 \end{array}$$

Substituting the value of $y = 38$ in equation (ii) above, $x + 38 = 114$

Therefore, $x = 114 - 38 = 76$

Hence, the missing frequencies are:

Against accident 1 : 76

Against accident 2 : 38

2.2.3 CHARACTERISTICS OF THE ARITHMETIC MEAN

Some of the important characteristics of the arithmetic mean are:

1. The sum of the deviations of the individual items from the arithmetic mean is always zero. This means I: $(x - \bar{x}) = 0$, where x is the value of an item and x is the arithmetic mean. Since the sum of the deviations in the positive direction is equal to the sum of the deviations in the negative direction, the arithmetic mean is regarded as a measure of central tendency.

2. The sum of the squared deviations of the individual items from the arithmetic mean is always minimum. In other words, the sum of the squared deviations taken from any value other than the arithmetic mean will be higher.

3. As the arithmetic mean is based on all the items in a series, a change in the value of any item will lead to a change in the value of the arithmetic mean.

4. In the case of highly skewed distribution, the arithmetic mean may get distorted on account of a few items with extreme values. In such a case, it may cease to be the representative characteristic of the distribution.

2.3 MEDIAN

Median is defined as the value of the middle item (or the mean of the values of the two middle items) when the data are arranged in an ascending or descending order of magnitude. Thus, in an ungrouped frequency distribution if the *n* values are arranged in ascending or descending order of magnitude, the median is the middle value if *n* is odd. When *n* is even, the median is the mean of the two middle values.

Suppose we have the following series:

15, 19,21,7, 10,33,25,18 and 5

We have to first arrange it in either ascending or descending order. These figures are arranged in an ascending order as follows:

5,7,10,15,18,19,21,25,33

Now as the series consists of odd number of items, to find out the value of the middle item, we use the formula

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

Where *n* is the number of items. In this case, *n* is 9, as such $\frac{n+1}{2}$ = 5, that is, the size of the 5th item is the median. This happens to be 18.

Suppose the series consists of one more items 23. We may, therefore, have to include 23 in the above series at an appropriate place, that is, between 21 and 25. Thus, the series is now 5, 7, 10, 15, 18, 19, and 21,23,25,33. Applying the above formula, the median is the size of 5.5^{th} item. Here, we have to take the average of the values of 5th and 6th item. This means an average of 18 and 19, which gives the median as 18.5.

It may be noted that the formula $\frac{n+1}{2}$ itself is not the formula for the median; it merely indicates the position of the median, namely, the number of items we have to count until we arrive at the item whose value is the median. In the case of the even number of items in the series, we identify the two items whose values have to be averaged to obtain the median. In the case of a grouped series, the median is calculated by linear interpolation with the help of the following formula:

$$\text{M} = l_1 \frac{l_2 + l_1}{f}(m - c)$$

Where M = the median

l_1 = the lower limit of the class in which the median lies

l_2 = the upper limit of the class in which the median lies

f = the frequency of the class in which the median lies

m = the middle item or $(n + 1)/2$th, where n stands for total number of

items

c = the cumulative frequency of the class preceding the one in which the median lies

Example 2.7:

Monthly Wages (Rs)	*No. of Workers*
800-1,000	18
1,000-1,200	25
1,200-1,400	30
1,400-1,600	34
1,600-1,800	26
1,800-2,000	10
Total	143

In order to calculate median in this case, we have to first provide cumulative frequency to the table. Thus, the table with the cumulative frequency is written as:

Monthly Wages	Frequency	Cumulative Frequency
800 -1,000	18	18
1,000 -1,200	25	43
1,200 -1,400	30	73
1,400 -1,600	34	107
1,600 -1,800	26	133
1.800 -2,000	10	143

$$\text{M} = l_1 \frac{l_2 + l_1}{f}(m - c)$$

$$\text{M} = \frac{n+1}{2} = \frac{143+1}{2} = 72$$

It means median lies in the class-interval Rs 1,200 - 1,400.

Now, $\text{M} = \mathit{1200} + \frac{1400 - 1200}{30}(72 - 43)$

$$= 1200 + \frac{200}{30}(29)$$

= Rs 1393.3

At this stage, let us introduce two other concepts viz. quartile and decile. To understand these, we should first know that the median belongs to a general class of statistical descriptions *called fractiles.* A fractile is a value below that lays a given fraction of a set of data. In the case of the median, this fraction is one-half (1/2). Likewise, a quartile has a fraction one-fourth (1/4). The three quartiles Q_1, Q_2 and Q_3 are such that 25 percent of the data fall below Q_1, 25 percent fall between Q_1 and Q_2, 25 percent fall between Q_2 and Q_3 and 25 percent fall above Q_3 It will be seen that Q_2 is the median. We can use the above formula for the calculation of quartiles as well. The only difference will be in the value of m. Let us calculate both Q_1 and Q_3 in respect of the table given in Example 2.7.

$$Q_1 \quad = \quad l_1 \frac{l_2 - l_1}{f}(m - c)$$

Here, m will be $= \frac{n+1}{4} = \frac{143+1}{4} = 36$

$$Q_1 = 1000 + \frac{1200-1000}{25}(36-18)$$

$$= 1000 + \frac{200}{25}(18)$$

= Rs. 1,144

In the case of Q_3, m will be 3 = $\frac{n+1}{4} = \frac{3 \times 144}{4}$ = 108

$$Q_1 = 1600 + \frac{1800-1600}{26}(108-107)$$

$$= 1600 + \frac{200}{26}(1)$$

Rs. 1,607.7 approx

In the same manner, we can calculate deciles (where the series is divided into 10 parts) and percentiles (where the series is divided into 100 parts). It may be noted that unlike arithmetic mean, median is not affected at all by extreme values, as it is a positional average. As such, median is particularly very useful when a distribution happens to be skewed. Another point that goes in favour of median is that it can be computed when a distribution has open-end classes. Yet, another merit of median is that when a distribution contains qualitative data, it is the only average that can be used. No other average is suitable in case of such a distribution. Let us take a couple of examples to illustrate what has been said in favour of median.

Example 2.8: Calculate the most suitable average for the following data:

Size of the Item	Below 50	50-100	100-150	150-200	200 and above
Frequency	*15*	*20*	*36*	*40*	*10*

Solution: Since the data have two open-end classes-one in the beginning (below 50) and the other at the end (200 and above), median should be the right choice as a measure of central tendency.

Table 2.6: Computation of Median

Size of Item	*Frequency*	*Cumulative Frequency*
Below 50	15	15
50-100	20	35
100-150	36	71
150-200	40	111
200 and above	10	121

Median is the size of $\frac{n+1}{2}$ th item

$= \frac{121+1}{2} = 61^{\text{st}}$ item

Now, 61^{st} item lies in the 100-150 class

Median = $\mathrm{l}_1 = l_1 \frac{l_2 - l_1}{f}(m - c)$

$= 100 + \frac{150-100}{36}(61-35)$

= 100 + 36.11 = 136.11 approx.

Example 2.9: The following data give the savings bank accounts balances of nine sample households selected in a survey. The figures are in rupees.

745 2,000 1,500 68,000 461 549 3750 1800 4795

(a) Find the mean and the median for these data; (b) Do these data contain an outlier? If so, exclude this value and recalculate the mean and median. Which of these summary measures has a greater change when an outlier is dropped?; (c) Which of these two summary measures is more appropriate for this series?

Solution:

Mean = Rs. $\dfrac{745 + 2{,}000 + 1{,}500 + 68{,}000 + 461 + 549 + 3{,}750 + 1{,}800 + 4{,}795}{9}$

= $\dfrac{\text{Rs } 83{,}600}{9}$ = Rs 9,289

Median = Size of $\dfrac{n + 1}{2}$ th item

= $\dfrac{9 + 1}{2}$ = 5$^{\text{th}}$ item

Arranging the data in an ascending order, we find that the median is Rs 1,800.

(b) An item of Rs 68,000 is excessively high. Such a figure is called an 'outlier'. We exclude this figure and recalculate both the mean and the median.

Mean = Rs. $\dfrac{83{,}600 - 68{,}000}{8}$

= Rs $\dfrac{15{,}600}{8}$ = Rs. 1,950

Median = Size of $\dfrac{n + 1}{2}$ th item

$= \dfrac{8 + 1}{2} = 4.5th$ item.

= Rs. $\dfrac{1{,}500 - 1{,}800}{2}$ = Rs. 1,650

It will be seen that the mean shows a far greater change than the median when the outlier is dropped from the calculations.

(c) As far as these data are concerned, the median will be a more appropriate measure than the mean.

Further, we can determine the median graphically as follows:

Example 2.10: Suppose we are given the following series:

Class interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	6	12	22	37	17	8	5

We are asked to draw both types of ogive from these data and to determine the median.

Solution:

First of all, we transform the given data into two cumulative frequency distributions, one based on 'less than' and another on 'more than' methods.

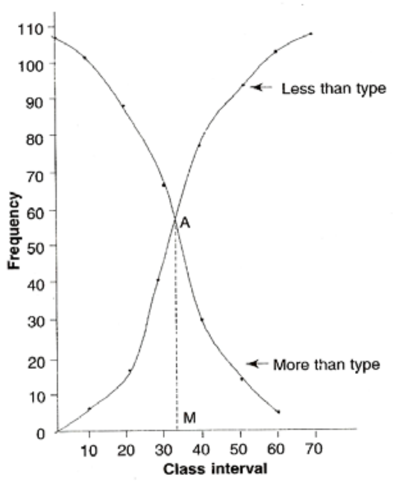
Table A

	Frequency
Less than 10	6
Less than 20	18
Less than 30	40
Less than 40	77
Less than 50	94
Less than 60	102
Less than 70	107

Table B

	Frequency
More than 0	107
More than 10	101
More than 20	89
More than 30	67
More than 40	30
More than 50	13
More than 60	5

It may be noted that the point of intersection of the two ogives gives the value of the median. From this point of intersection A, we draw a straight line to



meet the X-axis at M. Thus, from the point of origin to the point at M gives the value of the median, which comes to 34, approximately. If we calculate the median by applying the formula, then the answer comes to 33.8, or 34, approximately. It may be pointed out that even a single ogive can be used to determine the median. As we have determined the median graphically, so also we can find the values of quartiles, deciles or percentiles graphically. For example, to determine we have to take size of *{3(n* + 1)} /4 = 81st item. From this point on the Y-axis, we can draw a perpendicular to meet the 'less than' ogive from which another straight line is to be drawn to meet the X-axis. This point will give us the value of the upper quartile. In the same manner, other values of Q_1 and deciles and percentiles can be determined.

2.3.1 CHARACTERISTICS OF THE MEDIAN

1. Unlike the arithmetic mean, the median can be computed from open-ended distributions. This is because it is located in the median class-interval, which would not be an open-ended class.
2. The median can also be determined graphically whereas the arithmetic mean cannot be ascertained in this manner.
3. As it is not influenced by the extreme values, it is preferred in case of a distribution having extreme values.
4. In case of the qualitative data where the items are not counted or measured but are scored or ranked, it is the most appropriate measure of central tendency.

2.4 MODE

The mode is another measure of central tendency. It is the value at the point around which the items are most heavily concentrated. As an example, consider the following series: 8,9, 11, 15, 16, 12, 15,3, 7, 15

There are ten observations in the series wherein the figure 15 occurs maximum number of times three. The mode is therefore 15. The series given above is a discrete series; as such, the variable cannot be in fraction. If the series were continuous, we could say that the mode is approximately 15, without further computation.

In the case of grouped data, mode is determined by the following formula:

$$\text{Mode}= l_1 + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

Where, l_1 = the lower value of the class in which the mode lies

f_l = the frequency of the class in which the mode lies

f_o = the frequency of the class preceding the modal class

f_2 = the frequency of the class succeeding the modal class

i = the class-interval of the modal class

While applying the above formula, we should ensure that the class-intervals are uniform throughout. If the class-intervals are not uniform, then they should be made uniform on the assumption that the frequencies are evenly distributed throughout the class. In the case of inequal class-intervals, the application of the above formula will give misleading results.

Example 2.11: Let us take the following frequency distribution:

Class intervals (1)	*Frequency (2)*
30-40	4
40-50	6
50-60	8
60-70	12
70-80	9
80-90	7
90-100	4

We have to calculate the mode in respect of this series.

Solution: We can see from Column (2) of the table that the maximum frequency of 12 lies in the class-interval of 60-70. This suggests that the mode lies in this class-interval. Applying the formula given earlier, we get:

$$\text{Mode} = 60 + \frac{12 - 8}{12 - 8\,(12 - 8) + (12 - 9)} \times 10$$

$$= 60 + \frac{4}{4 + 3} \times 10$$

= 65.7 approx.

In several cases, just by inspection one can identify the class-interval in which the mode lies. One should see which the highest frequency is and then identify to which class-interval this frequency belongs. Having done this, the formula given for calculating the mode in a grouped frequency distribution can be applied.

At times, it is not possible to identify by inspection the class where the mode lies. In such cases, it becomes necessary to use the method of grouping. This method consists of two parts:

(i) **Preparation of a grouping table:** A grouping table has six columns, the first column showing the frequencies as given in the problem. Column 2 shows frequencies grouped in two's, starting from the top. Leaving the first frequency, column 3 shows frequencies grouped in two's. Column 4 shows the frequencies of the first three items, then second to fourth item and so on. Column 5 leaves the first frequency and groups the remaining items in three's. Column 6 leaves the first two frequencies and then groups the remaining in three's. Now, the maximum total in each column is marked and shown either in a circle or in a bold type.

(ii) **Preparation of an analysis table**: After having prepared a grouping table, an analysis table is prepared. On the left-hand side, provide the first column for column numbers and on the right-hand side the different possible values of mode. The highest values marked in the grouping table are shown here by a bar or by simply entering 1 in the relevant cell corresponding to the values they represent. The last row of this table will show the number of times a particular value has occurred in the grouping table. The highest value in the analysis table will indicate the class-interval in which the mode lies. The procedure of preparing both the grouping and analysis tables to locate the modal class will be clear by taking an example.

Example 2.12: The following table gives some frequency data:

Size of Item	Frequency
10-20	10
20-30	18
30-40	25
40-50	26
50-60	17
60-70	4

Solution:

Grouping Table

Size of item	1	2	3	4	5	6
10-20	10	28		53		
20-30	18		43		69	
30-40	25	51				68
40-50	26		43	47		
50-60	17	21				
60-70	4					

Analysis table

	Size of item				
Col. No.	10-20	20-30	30-40	40-50	50-60
1				1	
2			1	1	
3		1	1	1	1
4	1	1	1		
5		1	1	1	

6			1	1	1
Total	1	3	5	5	2

This is a bi-modal series as is evident from the analysis table, which shows that the two classes 30-40 and 40-50 have occurred five times each in the grouping. In such a situation, we may have to determine mode indirectly by applying the following formula:

Mode = 3 median - 2 mean

Median = Size of *(n* + l)/2th item, that is, 101/2 = 50.5th item. This lies in the class 30-40. Applying the formula for the median, as given earlier, we get

$$= \quad 30 + \frac{40 - 30}{25}(50.5 - 28)$$

$$= \quad 30 + 9 = 39$$

Now, arithmetic mean is to be calculated. This is shown in the following table.

Class- interval	Frequency	Mid- points	d	d' = d/10	fd'
10-20	10	15	-20	-2	-20
20-30	18	25	-10	-I	-18
30-40	25	35	0	0	0
40-50	26	45	10	1	26
50-60	17	55	20	2	34
60-70	4	65	30	3	12
Total	100				34

Deviation is taken from arbitrary mean = 35

$$\text{Mean} \quad = \quad A + \frac{\sum fd'}{n} \times i$$

$$= \quad 35 + \frac{34}{100} \times 10$$

$$= \quad 38.4$$

Mode = 3 median - 2 mean

= (3 x 39) - (2 x 38.4)

= 117 -76.8

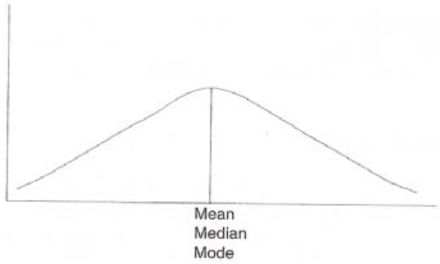
= 40.2

This formula, Mode = 3 Median-2 Mean, is an empirical formula only. And it can give only approximate results. As such, its frequent use should be avoided. However, when mode is ill defined or the series is bimodal (as is the case in the present example) it may be used.

2.5 RELATIONSHIPS OF THE MEAN, MEDIAN AND MODE

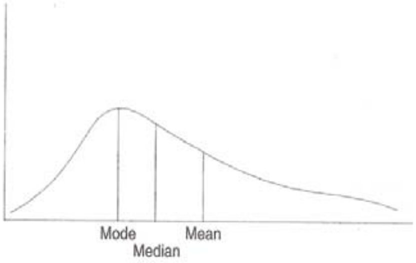
Having discussed mean, median and mode, we now turn to the relationship amongst these three measures of central tendency. We shall discuss the relationship assuming that there is a unimodal frequency distribution.

(i) When a distribution is symmetrical, the mean, median and mode are the same, as is shown below in the following figure.



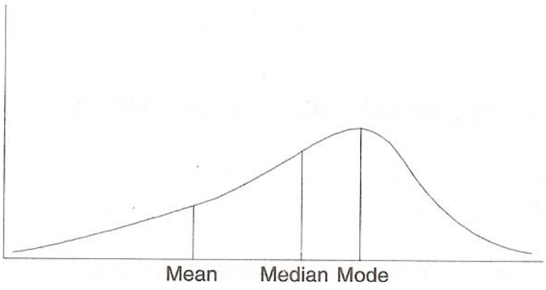
In case, a distribution is skewed to the right, then mean> median> mode. Generally, income distribution is skewed to the right where a large number of families have relatively low income and a small number of families have extremely high income. In such a case, the mean is pulled up by the extreme high incomes and the relation among these three measures is as shown in Fig. 6.3. Here, we find that mean> median> mode.

(ii) When a distribution is skewed to the left, then mode> median> mean. This is because here mean is pulled down below the median by extremely low values. This is



shown as in the figure.

(iii) Given the mean and median of a unimodal distribution, we can determine whether it is skewed to the right or left. When mean> median, it is skewed to the right; when median> mean, it is skewed to the left. It may be noted that the median is always in the middle between mean and mode.



2.6 THE BEST MEASURE OF CENTRAL TENDENCY

At this stage, one may ask as to which of these three measures of central tendency the best is. There is no simple answer to this question. It is because these three measures are based upon different concepts. The arithmetic mean is the sum of the values divided by the total number of observations in the series. The median is the value of the middle observation that divides the series into two equal parts. Mode is the value around which the observations tend to concentrate. As such, the use of a particular measure will largely depend on the purpose of the study and the nature of the data; For example, when we are interested in knowing the consumers preferences for different brands of television sets or different kinds of advertising, the choice should go in favour of mode. The use of mean and median would not be proper. However, the median can sometimes be used in the case of qualitative data when such data can be arranged in an ascending or descending order. Let us take another example. Suppose we invite applications for a certain vacancy in our company. A large number of candidates apply for that post. We are now interested to know as to which age or age group has the largest concentration of applicants. Here, obviously the mode will be the most appropriate choice. The arithmetic mean may not be appropriate as it may be influenced by some extreme values. However, the mean happens to be the most commonly used measure of central tendency as will be evident from the discussion in the subsequent chapters.

2.7 GEOMETRIC MEAN

Apart from the three measures of central tendency as discussed above, there are two other means that are used sometimes in business and economics. These are the geometric mean and the harmonic mean. The geometric mean is more important than the harmonic mean. We discuss below both these means. First, we take up the geometric mean. Geometric mean is defined at the *nth* root of the product of *n* observations of a distribution.

Symbolically, GM = $n\sqrt{x_1....x_2.....x_n...}$ If we have only two observations, say, 4 and 16 then GM = $\sqrt{4\times 16}=\sqrt{64}=8.$ Similarly, if there are three observations, then we have to calculate the cube root of the product of these three observations; and so on. When the number of items is large, it becomes extremely difficult to multiply the numbers and to calculate the root. To simplify calculations, logarithms are used.

Example 2.13: If we have to find out the geometric mean of 2, 4 and 8, then we find

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

$$= \frac{Log2 + Log4 + Log8}{3}$$

$$= \frac{0.3010 + 0.6021 + 0.9031}{3}$$

$$= \frac{1.8062}{3} = 0.60206$$

$$\text{GM} = \text{Antilog } 0.60206$$

$$= 4$$

When the data are given in the form of a frequency distribution, then the geometric mean can be obtained by the formula:

$$\text{Log GM} = \frac{f_1 .\log x_l + f_2 .\log x_2 + ... + f_n . \log x_n}{f_1 + f_2 + ...fn}$$

$$= \frac{\sum f.\log x}{f_1 + f_2 + ...fn}$$

Then, GM = Antilog *n*

The geometric mean is most suitable in the following three cases:

1. Averaging rates of change.
2. The compound interest formula.
3. Discounting, capitalization.

Example 2.14: A person has invested Rs 5,000 in the stock market. At the end of the first year the amount has grown to Rs 6,250; he has had a 25 percent profit. If at the end of the second year his principal has grown to Rs 8,750, the rate of increase is 40 percent for the year. What is the average rate of increase of his investment during the two years?

Solution:

GM = $\sqrt{1.25 \times 1.40} = \sqrt{1.75.}$ = 1.323

The average rate of increase in the value of investment is therefore 1.323 - 1 = 0.323, which if multiplied by 100, gives the rate of increase as 32.3 percent.

Example 2.15: We can also derive a compound interest formula from the above set of data. This is shown below:

Solution: Now, 1.25 x 1.40 = 1.75. This can be written as 1.75 = $(1 + 0.323)^2$.

Let P_2 = 1.75, P_0 = 1, and r = 0.323, then the above equation can be written as $P_2 = (1 + r)^2$ or $P_2 = P_0 (1 + r)^2$.

Where *P2* is the value of investment at the end of the second year, P_0 is the initial investment and *r* is the rate of increase in the two years. This, in fact, is the familiar compound interest formula. This can be written in a generalised form as $P_n = P_0(1 + r)^n$. In our case *Po* is Rs 5,000 and the rate of increase in investment is 32.3 percent. Let us apply this formula to ascertain the value of *Pn, that* is, investment at the end of the second year.

$P_n = 5{,}000\ (1 + 0.323)^2$

= 5,000 x 1.75

= Rs 8,750

It may be noted that in the above example, if the arithmetic mean is used, the resultant figure will be wrong. In this case, the average rate for the two years is $\frac{25+40}{2}$ percent per year, which comes to 32.5. Applying this rate, we get $P_n = \frac{165}{100}$ x 5,000

= Rs 8,250

This is obviously wrong, as the figure should have been Rs 8,750.

Example 2.16: An economy has grown at 5 percent in the first year, 6 percent in the second year, 4.5 percent in the third year, 3 percent in the fourth year and 7.5 percent in the fifth year. What is the average rate of growth of the economy during the five years?

Solution:

Year	*Rate of Growth (percent)*	*Value at the end of the Year x (in Rs)*	*Log x*
1	5	105	2.02119
2	6	106	2.02531
3	4.5	104.5	2.01912
4	3	103	2.01284
5	7.5	107.5	2.03141
			$\sum$ log X = 10.10987

$$\text{GM} = \text{Antilog}\left(\frac{\sum \log x}{n}\right)$$

$$= \text{Antilog}\left(\frac{10.10987}{5}\right)$$

= Antilog 2.021974

= 105.19

Hence, the average rate of growth during the five-year period is 105.19 - 100 = 5.19 percent per annum. In case of a simple arithmetic average, the corresponding rate of growth would have been 5.2 percent per annum.

2.7.1 DISCOUNTING

The compound interest formula given above was

$P_n=P_0(1+r)^n$ This can be written as $P_0 = \dfrac{P_n}{(1+r)^n}$

This may be expressed as follows:

If the future income is P_n rupees and the present rate of interest is 100 r percent, then the present value of P n rupees will be P_0 rupees. For example, if we have a machine that has a life of 20 years and is expected to yield a net income of Rs 50,000 per year, and at the end of 20 years it will be obsolete and cannot be used, then the machine's present value is

$$\frac{50{,}000}{(1+r)^n}+\frac{50{,}000}{(1+r)^2}+\frac{50{,}000}{(1+r)^3}+\ldots\ldots\ldots\ldots\ldots\ldots\frac{50{,}000}{(1+r)^{20}}$$

This process of ascertaining the present value of future income by using the interest rate is known as discounting.

In conclusion, it may be said that when there are extreme values in a series, geometric mean should be used as it is much less affected by such values. The arithmetic mean in such cases will give misleading results.

Before we close our discussion on the geometric mean, we should be aware of its advantages and limitations.

2.7.2 ADVANTAGES OF G. M.

1. Geometric mean is based on each and every observation in the data set.
2. It is rigidly defined.
3. It is more suitable while averaging ratios and percentages as also in calculating growth rates.
4. As compared to the arithmetic mean, it gives more weight to small values and less weight to large values. As a result of this characteristic of the geometric mean, it is generally less than the arithmetic mean. At times it may be equal to the arithmetic mean.
5. It is capable of algebraic manipulation. If the geometric mean has two or more series is known along with their respective frequencies. Then a combined geometric mean can be calculated by using the logarithms.

2.7.3 LIMITATIONS OF G.M.

1. As compared to the arithmetic mean, geometric mean is difficult to understand.
2. Both computation of the geometric mean and its interpretation are rather difficult.
3. When there is a negative item in a series or one or more observations have zero value, then the geometric mean cannot be calculated.

In view of the limitations mentioned above, the geometric mean is not frequently used.

2.8 HARMONIC MEAN

The harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocals of individual observations. Symbolically,

$$\text{HM}=\frac{n}{1/\text{x}_1+1/\text{x}_2+1/\text{x}_3+\ldots+1/\text{x}_\text{n}}=\text{Re}\,ciprocal\frac{\sum 1/x}{n}$$

The calculation of harmonic mean becomes very tedious when a distribution has a large number of observations. In the case of grouped data, the harmonic mean is calculated by using the following formula:

$$\text{HM} = \text{Reciprocal of } \sum_{i-1}^{n}\left(f_i \times \frac{1}{x_i}\right)$$

or

$$\frac{n}{\sum_{i-1}^{n}\left(f_i \times \frac{1}{x_i}\right)}$$

Where *n* is the total number of observations.

Here, each reciprocal of the original figure is weighted by the corresponding frequency *(f)*.

The main **advantage** of the harmonic mean is that it is based on all observations in a distribution and is amenable to further algebraic treatment. When we desire to give greater weight to smaller observations and less weight to the larger observations, then the use of harmonic mean will be more suitable. As against these advantages, there are certain limitations of the harmonic mean. First, it is difficult to understand as well as difficult to compute. Second, it cannot be calculated if any of the observations is zero or negative. Third, it is only a summary figure, which may not be an actual observation in the distribution.

It is worth noting that the harmonic mean is always lower than the geometric mean, which is lower than the arithmetic mean. This is because the harmonic mean assigns lesser importance to higher values. Since the harmonic mean is based on reciprocals, it becomes clear that as reciprocals of higher values are lower than those of lower values, it is a lower average than the arithmetic mean as well as the geometric mean.

Example 2.17: Suppose we have three observations 4, 8 and 16. We are required to calculate the harmonic mean. Reciprocals of 4,8 and 16 are: $\frac{1}{4}, \frac{1}{8}, \frac{1}{16}$ respectively

Since HM = $\dfrac{n}{1/\text{x}_1 + 1/\text{x}_2 + 1/\text{x}_3}$

$$= \frac{3}{1/4 + 1/8 + 1/16}$$

$$= \frac{3}{0.25 + 0.125 + 0.0625}$$

= 6.857 approx.

Example 2.18: Consider the following series:

Class-interval	2-4	4-6	6-8	8-10
Frequency	20	40	30	10

Solution:

Let us set up the table as follows:

Class-interval	Mid-value	Frequency	Reciprocal of MV	f x 1/x
2-4	3	20	0.3333	6.6660
4-6	5	40	0.2000	8.0000
6-8	7	30	0.1429	4.2870
8-10	9	10	0.1111	1.1111
			Total	20.0641

$$= \frac{\sum_{i-1}^{n}\left(f_i \times \frac{1}{x_i}\right)}{n}$$

$$= \frac{100}{20.0641} = 4.984 \text{ approx.}$$

Example 2.19: In a small company, two typists are employed. Typist A types one page in ten minutes while typist B takes twenty minutes for the same. (i) Both are asked to type 10 pages. What is the average time taken for typing one page? (ii) Both are asked to type for one hour. What is the average time taken by them for typing one page?

Solution: Here Q-(i) is on arithmetic mean while Q-(ii) is on harmonic mean.

(i) $$\text{M} = \frac{(10\times10)+(20\times20)(\min utes)}{10\times2(pages)}$$

$$= \quad 15 \text{ minutes}$$

$$\text{HM} \quad = \quad \frac{60\times(\min utes)}{60/10+60/20(pages)}$$

$$= \quad \frac{120}{\frac{120+60}{20}} = \frac{40}{3} = 13\min utes \text{ and 20 seconds.}$$

Example 2.20: It takes ship A 10 days to cross the Pacific Ocean; ship B takes 15 days and ship C takes 20 days. (i) What is the average number of days taken by a ship to cross the Pacific Ocean? (ii) What is the average number of days taken by a cargo to cross the Pacific Ocean when the ships are hired for 60 days?

Solution: Here again Q-(i) pertains to simple arithmetic mean while Q-(ii) is concerned with the harmonic mean.

(i) $$\text{M} \quad = \quad \frac{10+15+20}{3} = 15 \text{ days}$$

(ii) $$\text{HM} \quad = \quad \frac{60\times3(days)\_}{60/10+60/15+60/20}$$

$$= \quad \frac{180}{\frac{360+240+180}{60}}$$

= 13.8 days approx.

2.9 QUADRATIC MEAN

We have seen earlier that the geometric mean is the antilogarithm of the arithmetic mean of the logarithms, and the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. Likewise, the quadratic mean (Q) is the square root of the arithmetic mean of the squares. Symbolically,

$$Q = \sqrt{\frac{x_1^2 + x_2^2 + \ldots\ldots +_n^2}{n}}$$

Instead of using original values, the quadratic mean can be used while averaging deviations when the standard deviation is to be calculated. This will be used in the next chapter on dispersion.

2.9.1 Relative Position of Different Means

The relative position of different means will always be:

$Q > \bar{x} > G > H$ provided that all the individual observations in a series are positive and all of them are not the same.

2.9.2 Composite Average or Average of Means

Sometimes, we may have to calculate an average of several averages. In such cases, we should use the same method of averaging that was employed in calculating the original averages. Thus, we should calculate the arithmetic mean of several values of *x,* the geometric mean of several values of GM, and the harmonic mean of several values of HM. It will be wrong if we use some other average in averaging of means.

OBJECTIVE: The objective of the present lesson is to impart the knowledge of measures of dispersion and skewness and to enable the students to distinguish between average, dispersion, skewness, moments and kurtosis.

STRUCTURE:

INTRODUCTION

In the previous chapter, we have explained the measures of central tendency. It may be noted that these measures do not indicate the extent of dispersion or variability in a distribution. The dispersion or variability provides us one more step in increasing our understanding of the pattern of the data. Further, a high degree of uniformity (i.e. low degree of dispersion) is a desirable quality. If in a business there is a high degree of variability in the raw material, then it could not find mass production economical.

Suppose an investor is looking for a suitable equity share for investment. While examining the movement of share prices, he should avoid those shares that are highly fluctuating-having sometimes very high prices and at other times going very low. Such extreme fluctuations mean that there is a high risk in the investment in shares. The investor should, therefore, prefer those shares where risk is not so high.

MEANING AND DEFINITIONS OF DISPERSION

The various measures of central value give us one single figure that represents the entire data. But the average alone cannot adequately describe a set of observations, unless all the observations are the same. It is necessary to describe the variability or dispersion of the observations. In two or more distributions the central value may be the same but still there can be wide disparities in the formation of distribution. Measures of dispersion help us in studying this important characteristic of a distribution.

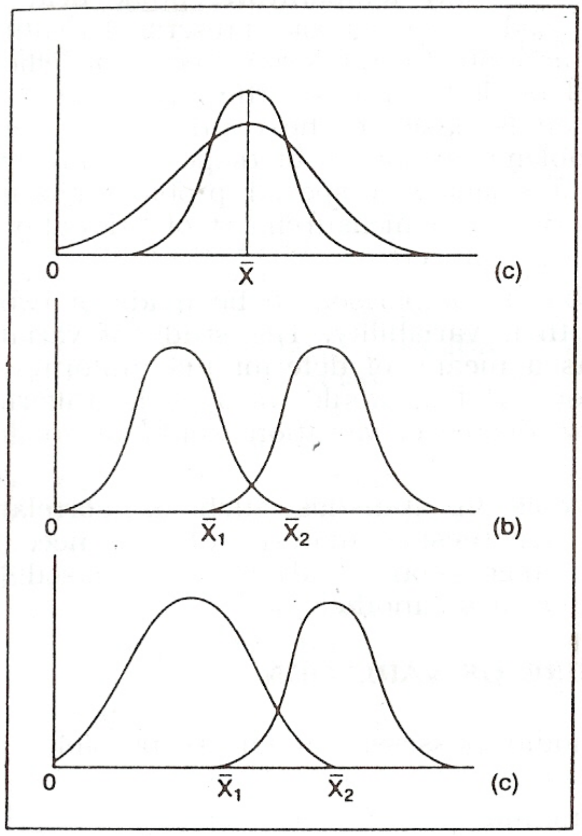
Some important definitions of dispersion are given below:

1. "Dispersion is the measure of the variation of the items." -A.L. Bowley
2. "The degree to which numerical data tend to spread about an average value is called the variation of dispersion of the data." -Spiegel
3. Dispersion or spread is the degree of the scatter or variation of the variable about a central value." -Brooks & Dick
4. "The measurement of the scatterness of the mass of figures in a series about an average is called measure of variation or dispersion." -Simpson & Kajka

It is clear from above that dispersion (also known as scatter, spread or variation) measures the extent to which the items vary from some central value. Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of the second order. An average is more meaningful when it is examined in the light of dispersion. For example, if the average wage of the workers of factory A is Rs. 3885 and that of factory B Rs. 3900, we cannot necessarily conclude that the workers of factory B are better off because in factory B there may be much greater dispersion in the distribution of wages. The study of dispersion is of great significance in practice as could well be appreciated from the following example:

	Series A	**Series B**	**Series C**
	100	100	1
	100	105	489
	100	102	2
	100	103	3
	100	90	5
Total	500	500	500
$\bar{x}$	100	100	100

Since arithmetic mean is the same in all three series, one is likely to conclude that these series are alike in nature. But a close examination shall reveal that distributions differ widely from one another. In series A, (In Box-3.1) each and every item is perfectly represented by the arithmetic mean or in other words none of the items of series A deviates from the



arithmetic mean and hence there is no dispersion. In series B, only one item is perfectly represented by the arithmetic mean and the other items vary but the variation is very small as compared to series C. In series C. not a single item is represented by the arithmetic mean and the items vary widely from one another. In series C, dispersion is much greater compared to series B. Similarly, we may have two groups of labourers with the same mean salary and yet their distributions may differ widely. The mean salary may not be so important a characteristic as the variation of the items from the mean. To the student of social affairs the mean income is not so vitally important as to know how this income is distributed. Are a large number receiving the mean income or are there a few with enormous incomes and millions with incomes far below the mean? The three figures given in Box 3.1 represent frequency distributions with some of the characteristics. The two curves in diagram (a) represent two distractions with the same mean $\overline{X}$, but with different dispersions. The two curves in (b) represent two distributions with the same dispersion but with unequal means $\overline{X}_1$ and $\overline{X}_2$, (c) represents two distributions with unequal dispersion. The measures of central tendency are, therefore insufficient. They must be supported and supplemented with other measures.

In the present chapter, we shall be especially concerned with the measures of variability or spread or dispersion. A measure of variation or dispersion is one that measures the extent to which there are differences between individual observation and some central or average value. In measuring variation we shall be interested in the amount of the variation or its degree but not in the direction. For example, a measure of 6 inches below the mean has just as much dispersion as a measure of six inches above the mean.

Literally meaning of dispersion is ‘scatteredness’. Average or the measures of central tendency gives us an idea of the concentration of the observations about the central part of the distribution. If we know the average alone, we cannot form a complete idea about the distribution. But with the help of dispersion, we have an idea about homogeneity or heterogeneity of the distribution.

3.3 SIGNIFICANCE AND PROPERTIES OF MEASURING VARIATION

Measures of variation are needed for four basic purposes:

1. Measures of variation point out as to how far an average is representative of the mass. When dispersion is small, the average is a typical value in the sense that it closely represents the individual value and it is reliable in the sense that it is a good estimate of the average in the corresponding universe. On the other hand, when dispersion is large, the average is not so typical, and unless the sample is very large, the average may be quite unreliable.

2. Another purpose of measuring dispersion is to determine nature and cause of variation in order to control the variation itself. In matters of health variations in body temperature, pulse beat and blood pressure are the basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production efficient operation requires control of quality variation the causes of which are sought through inspection is basic to the control of causes of variation. In social sciences a special problem requiring the measurement of variability is the measurement of "inequality" of the distribution of income or wealth etc.

3. Measures of dispersion enable a comparison to be made of two or more series with regard to their variability. The study of variation may also be looked

upon as a means of determining uniformity of consistency. A high degree of variation would mean little uniformity or consistency whereas a low degree of variation would mean great uniformity or consistency.

4. Many powerful analytical tools in statistics such as correlation analysis. the testing of hypothesis, analysis of variance, the statistical quality control, regression analysis is based on measures of variation of one kind or another.

A good measure of dispersion should possess the following properties

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be rigidly defined.
4. It should be based on each and every item of the distribution.
5. It should be amenable to further algebraic treatment.
6. It should have sampling stability.
7. Extreme items should not unduly affect it.

3.4 MEAURES OF DISPERSION

There are five measures of dispersion: Range, Inter-quartile range or Quartile Deviation, Mean deviation, Standard Deviation, and Lorenz curve. Among them, the first four are mathematical methods and the last one is the graphical method. These are discussed in the ensuing paragraphs with suitable examples.

3.5 RANGE

The simplest measure of dispersion is the range, which is the difference between the maximum value and the minimum value of data.

Example 3.1: Find the range for the following three sets of data:

Set 1:	05	15	15	05	15	05	15	15	15	15
Set 2:	8	7	15	11	12	5	13	11	15	9

Set 3:	5	5	5	5	5	5	5	5	5	5

Solution: In each of these three sets, the highest number is 15 and the lowest number is 5. Since the range is the difference between the maximum value and the minimum value of the data, it is 10 in each case. But the range fails to give any idea about the dispersal or spread of the series between the highest and the lowest value. This becomes evident from the above data.

In a frequency distribution, range is calculated by taking the difference between the upper limit of the highest class and the lower limit of the lowest class.

Example 3.2: Find the range for the following frequency distribution:

Size of Item	Frequency
20- 40	7
40- 60	11
60- 80	30
80-100	17
100-120	5
Total	**70**

Solution: Here, the upper limit of the highest class is 120 and the lower limit of the lowest class is 20. Hence, the range is 120 - 20 = 100. Note that the range is not influenced by the frequencies. Symbolically, the range is calculated b the formula L - S, where L is the largest value and S is the smallest value in a distribution. The coefficient of range is calculated by the formula: (L-S)/ (L+S). This is the relative measure. The coefficient of the range in respect of the earlier example having three sets of data is: 0.5.The coefficient of range is more appropriate for purposes of comparison as will be evident from the following example:

Example 3.3: Calculate the coefficient of range separately for the two sets of data given below:

Set 1	8	10	20	9	15	10	13	28
Set 2	30	35	42	50	32	49	39	33

Solution: It can be seen that the range in both the sets of data is the same:

Set 1 28 - 8 = 20

Set 2 50 - 30 = 20

Coefficient of range in Set 1 is:

$$\frac{28-8}{28+8} = 0.55$$

Coefficient of range in set 2 is:

$$\frac{50-30}{50+30} = 0.25$$

3.5.1 LIMITATIONS OF RANGE

There are some limitations of range, which are as follows:

1. It is based only on two items and does not cover all the items in a distribution.
2. It is subject to wide fluctuations from sample to sample based on the same population.
3. It fails to give any idea about the pattern of distribution. This was evident from the data given in Examples 1 and 3.
4. Finally, in the case of open-ended distributions, it is not possible to compute the range.

Despite these limitations of the range, it is mainly used in situations where one wants to quickly have some idea of the variability or' a set of data. When the sample size is very small, the range is considered quite adequate measure of the variability. Thus, it is widely used in quality control where a continuous check on the variability of raw materials or finished products is needed. The range is also a suitable measure in weather forecast. The meteorological department uses the range by giving the maximum and the minimum temperatures. This information is quite useful to the common man, as he can know the extent of possible variation in the temperature on a particular day.

3.6 INTERQUARTILE RANGE OR QUARTILE DEVIATION

The interquartile range or the quartile deviation is a better measure of variation in a distribution than the range. Here, avoiding the 25 percent of the distribution at both the ends uses the middle 50 percent of the distribution. In other words, the interquartile range denotes the difference between the third quartile and the first quartile.

Symbolically, interquartile range = $Q_3 - Q_l$

Many times the interquartile range is reduced in the form of semi-interquartile range or quartile deviation as shown below:

Semi interquartile range or Quartile deviation = $(Q_3 – Q_l)/2$

When quartile deviation is small, it means that there is a small deviation in the central 50 percent items. In contrast, if the quartile deviation is high, it shows that the central 50 percent items have a large variation. It may be noted that in a symmetrical distribution, the two quartiles, that is, Q3 and QI are equidistant from the median. Symbolically,

$M-Q_I = Q_3-M$

However, this is seldom the case as most of the business and economic data are asymmetrical. But, one can assume that approximately 50 percent of the observations are contained in the interquartile range. It may be noted that interquartile range or the quartile deviation is an absolute measure of dispersion. It can be changed into a relative measure of dispersion as follows:

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

The computation of a quartile deviation is very simple, involving the computation of upper and lower quartiles. As the computation of the two quartiles has already been explained in the preceding chapter, it is not attempted here.

3.6.1 **MERITS OF QUARTILE DEVIATION**

The following merits are entertained by quartile deviation:

1. As compared to range, it is considered a superior measure of dispersion.
2. In the case of open-ended distribution, it is quite suitable.
3. Since it is not influenced by the extreme values in a distribution, it is particularly suitable in highly skewed or erratic distributions.

3.6.2 LIMITATIONS OF QUARTILE DEVIATION

1. Like the range, it fails to cover all the items in a distribution.
2. It is not amenable to mathematical manipulation.
3. It varies widely from sample to sample based on the same population.
4. Since it is a positional average, it is not considered as a measure of dispersion. It merely shows a distance on scale and not a scatter around an average.

In view of the above-mentioned limitations, the interquartile range or the quartile deviation has a limited practical utility.

3.7 MEAN DEVIATION

The mean deviation is also known as the average deviation. As the name implies, it is the average of absolute amounts by which the individual items deviate from the mean. Since the positive deviations from the mean are equal to the negative deviations, while computing the mean deviation, we ignore positive and negative signs. Symbolically,

$\text{MD} = \frac{\sum |x|}{n}$ Where MD = mean deviation, |x| = deviation of an item from the mean ignoring positive and negative signs, *n* = the total number of observations.

Example 3.4:

Size of Item	Frequency
2-4	20
4-6	40
6-8	30
8-10	10

Solution:

Size of Item	Mid-points (m)	Frequency (f)	fm	d from $\overline{x}$	f |d|
2-4	3	20	60	-2.6	52
4-6	5	40	200	-0.6	24
6-8	7	30	210	1.4	42
8-10	9	10	90	3.4	34
	Total	**100**	**560**		**152**

$$\overline{x} = \frac{\sum fm}{n} = \frac{560}{100} = 5.6$$

$$\text{MD } (\overline{x}) = \frac{\sum f\ |d\ |}{n} = \frac{152}{100} = 1.52$$

3.7.1 MERITS OF MEAN DEVIATION

1. A major advantage of mean deviation is that it is simple to understand and easy to calculate.
2. It takes into consideration each and every item in the distribution. As a result, a change in the value of any item will have its effect on the magnitude of mean deviation.
3. The values of extreme items have less effect on the value of the mean deviation.
4. As deviations are taken from a central value, it is possible to have meaningful comparisons of the formation of different distributions.

3.7.2 LIMITATIONS OF MEAN DEVIATION

1. It is not capable of further algebraic treatment.

2. At times it may fail to give accurate results. The mean deviation gives best results when deviations are taken from the median instead of from the mean. But in a series, which has wide variations in the items, median is not a satisfactory measure.

3. Strictly on mathematical considerations, the method is wrong as it ignores the algebraic signs when the deviations are taken from the mean.

In view of these limitations, it is seldom used in business studies. A better measure known as the standard deviation is more frequently used.

3.8 STANDARD DEVIATION

The standard deviation is similar to the mean deviation in that here too the deviations are measured from the mean. At the same time, the standard deviation is preferred to the mean deviation or the quartile deviation or the range because it has desirable mathematical properties.

Before defining the concept of the standard deviation, we introduce another concept viz. variance.

Example 3.5:

X	X-μ	$(X-\mu)^2$
20	20-18=12	4
15	15-18= -3	9
19	19-18 = 1	1
24	24-18 = 6	36
16	16-18 = -2	4
14	14-18 = -4	16
108	**Total**	**70**

Solution:

$$\text{Mean} = \frac{108}{6} = 18$$

The second column shows the deviations from the mean. The third or the last column shows the squared deviations, the sum of which is 70. The arithmetic mean of the squared deviations is:

$$\frac{\sum (x-\mu)^2}{N} = 70/6 = 11.67 \text{ approx.}$$

This mean of the squared deviations is known as the variance. It may be noted that this variance is described by different terms that are used interchangeably: the variance of the distribution X; the variance of X; the variance of the distribution; and just simply, the variance.

Symbolically, Var (X) = $\frac{\sum (x-\mu)^2}{N}$

It is also written as $\sigma^2 = \frac{\sum (x_i-\mu)^2}{N}$

Where σ^2 (called sigma squared) is used to denote the variance.

Although the variance is a measure of dispersion, the unit of its measurement is (points). If a distribution relates to income of families then the variance is $(\text{Rs})^2$ and not rupees. Similarly, if another distribution pertains to marks of students, then the unit of variance is $(\text{marks})^2$. To overcome this inadequacy, the square root of variance is taken, which yields a better measure of dispersion known as the standard deviation. Taking our earlier example of individual observations, we take the square root of the variance

SD or $\sigma = \sqrt{Variance} = \sqrt{11.67} = 3.42$ points

Symbolically, $\sigma = \sqrt{\frac{\sum (x_i-\mu)^2}{N}}$

In applied Statistics, the standard deviation is more frequently used than the variance. This can also be written as:

$$\sigma = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}}{N}}$$

We use this formula to calculate the standard deviation from the individual observations given earlier.

Example 7.6:

X	X^2
20	400
15	225
19	361
24	576
16	256
14	196
108	2014

Solution:

$\sum x_i^2 = 2014$ $\quad \sum \mathrm{x}_i = 108 \quad$ N = 6

$$\sigma = \sqrt{\frac{2014 - \frac{(108)^2}{6}}{6}} \quad \text{Or,} \quad \sigma = \sqrt{\frac{2014 - \frac{11664}{6}}{6}}$$

$$\sigma = \sqrt{\frac{\frac{12084 - 11664}{6}}{6}} \quad \text{Or,} \quad \sigma = \sqrt{\frac{\frac{420}{6}}{6}}$$

$$\sigma = \sqrt{\frac{70}{6}} \quad \text{Or,} \quad \sigma = \sqrt{11.67}$$

$\sigma =$ 3.42

Example 3.7:

The following distribution relating to marks obtained by students in an examination:

Marks	Number of Students
0- 10	1
10- 20	3
20- 30	6
30- 40	10
40- 50	12
50- 60	11

60- 70	6
70- 80	3
80- 90	2
90-100	1

Solution:

Marks	**Frequency (f)**	**Mid-points**	**Deviations (d)/10=d'**	**Fd'**	**fd'^{2}**
0- 10	1	5	-5	-5	25
10- 20	3	15	-4	-12	48
20- 30	6	25	-3	-18	54
30- 40	10	35	-2	-20	40
40- 50	12	45	-1	-12	12
50- 60	11	55	0	0	0
60- 70	6	65	1	6	6
70- 80	3	75	2	6	12
80- 90	2	85	3	6	18
90-100	1	95	4	4	16
Total	**55**		**Total**	**-45**	**231**

In the case of frequency distribution where the individual values are not known, we use the midpoints of the class intervals. Thus, the formula used for calculating the standard deviation is as given below:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{K} fi(m_i - \mu)^2}{N}}$$

Where m$_i$ is the mid-point of the class intervals μ is the mean of the distribution, *fi* is the frequency of each class; N is the total number of frequency and K is the number of classes. This formula requires that the mean μ be calculated and that deviations (m$_i$ - μ) be obtained for each class. To avoid this inconvenience, the above formula can be modified as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{K} fid_i^2 \left(\sum_{i=1}^{K} fd_i \right)}{N}}$$

Where C is the class interval: f_i is the frequency of the *i*th class and d_i is the deviation of the of item from an assumed origin; and N is the total number of observations.

Applying this formula for the table given earlier,

$$\sigma = 10\sqrt{\frac{231}{55} - \left(\frac{-45}{55}\right)^2}$$

$$=10\sqrt{4.2-0.669421}$$

$$=18.8 \text{ marks}$$

When it becomes clear that the actual mean would turn out to be in fraction, calculating deviations from the mean would be too cumbersome. In such cases, an assumed mean is used and the deviations from it are calculated. While mid-point of any class can be taken as an assumed mean, it is advisable to choose the mid-point of that class that would make calculations least cumbersome. Guided by this consideration, in Example 3.7 we have decided to choose 55 as the mid-point and, accordingly, deviations have been taken from it. It will be seen from the calculations that they are considerably simplified.

3.8.1 USES OF THE STANDARD DEVIATION

The standard deviation is a frequently used measure of dispersion. It enables us to determine as to how far individual items in a distribution deviate from its mean. In a symmetrical, bell-shaped curve:

(i) About 68 percent of the values in the population fall within: ± 1 standard deviation from the mean.

(ii) About 95 percent of the values will fall within ±2 standard deviations from the mean.

(iii) About 99 percent of the values will fall within ± 3 standard deviations from the mean.

The standard deviation is an absolute measure of dispersion as it measures variation in the same units as the original data. As such, it cannot be a suitable measure while comparing two or more distributions. For this purpose, we should use a relative measure of dispersion. One such measure of relative dispersion is the coefficient of variation, which relates the standard deviation and the mean such that the standard deviation is expressed as a percentage of mean. Thus, the specific unit in which the standard deviation is measured is done away with and the new unit becomes percent.

Symbolically, CV (coefficient of variation) $= \frac{\sigma}{\mu} \text{x} 100$

Example 3.8: In a small business firm, two typists are employed-typist A and typist B. Typist A types out, on an average, 30 pages per day with a standard deviation of 6. Typist B, on an average, types out 45 pages with a standard deviation of 10. Which typist shows greater consistency in his output?

Solution: Coefficient of variation for $A = \frac{\sigma}{\mu} \text{x} 100$

Or $A = \frac{6}{30} \text{x} 100$

Or 20% and

Coefficient of variation for $B = \frac{\sigma}{\mu} \text{x} 100$

$B = \frac{10}{45} \text{x} 100$

or 22.2 %

These calculations clearly indicate that although typist B types out more pages, there is a greater variation in his output as compared to that of typist A. We can say this in a different way: Though typist A's daily output is much less, he is more consistent than typist B. The usefulness of the coefficient of variation becomes clear in comparing two groups of data having different means, as has been the case in the above example.

3.8.2 STANDARDISED VARIABLE, STANDARD SCORES

The variable $Z = (x - \bar{x})/s$ or $(x - \mu)/\mu$, which measures the deviation from the mean in units of the standard deviation, is called a standardised variable. Since both the numerator and the denominator are in the same units, a standardised variable is independent of units used. If deviations from the mean are given in units of the standard deviation, they are said to be expressed in standard units or standard scores.

Through this concept of standardised variable, proper comparisons can be made between individual observations belonging to two different distributions whose compositions differ.

Example 3.9: A student has scored 68 marks in Statistics for which the average marks were 60 and the standard deviation was 10. In the paper on Marketing, he scored 74 marks for which the average marks were 68 and the standard deviation was 15. In which paper, Statistics or Marketing, was his relative standing higher?

Solution: The standardised variable Z = (x - $\bar{x}$) ÷ s measures the deviation of x from the mean x in terms of standard deviation s. For Statistics, Z = (68 - 60) ÷ 10 = 0.8

For Marketing, Z = (74 - 68) ÷ 15 = 0.4

Since the standard score is 0.8 in Statistics as compared to 0.4 in Marketing, his relative standing was higher in Statistics.

Example 3.10: Convert the set of numbers 6, 7, 5, 10 and 12 into standard scores:

Solution:

X	X^2
6	36
7	49
5	25
10	100
12	144
$\sum X = 40$	$\sum X^2 = 354$

$$\bar{x} = \sum x \div N = 40 \div 5 = 8$$

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum X)^2}{N}}{N}} \quad \text{or,} \quad \sigma = \sqrt{\frac{354 - \frac{(40)^2}{5}}{5}}$$

$$= \sqrt{\frac{354 - 320}{5}} = 2.61 \text{ approx.}$$

$$Z = \frac{x - \bar{x}}{\sigma} = \frac{6-8}{2.61} = \text{ -0.77 (Standard score)}$$

Applying this formula to other values:

(i) $\frac{7-8}{2.61}$ = -0.38

(ii) $\frac{5-8}{2.61}$ = -1.15

(iii) $\frac{10-8}{2.61}$ = 0.77

(iv) $\frac{12-8}{2.61}$ = 1.53

Thus the standard scores for 6,7,5,10 and 12 are -0.77, -0.38, -1.15, 0.77 and 1.53, respectively.

3.9 LORENZ CURVE

This measure of dispersion is graphical. It is known as the Lorenz curve named after Dr. Max Lorenz. It is generally used to show the extent of concentration of income and wealth. The steps involved in plotting the Lorenz curve are:

1. Convert a frequency distribution into a cumulative frequency table.
2. Calculate percentage for each item taking the total equal to 100.
3. Choose a suitable scale and plot the cumulative percentages of the persons and income. Use the horizontal axis of X to depict percentages of persons and the vertical axis of Y to depict percent ages of income.
4. Show the line of equal distribution, which will join 0 of X-axis with 100 of Y-axis.
5. The curve obtained in (3) above can now be compared with the straight line of equal distribution obtained in (4) above. If the Lorenz curve is close to the line of equal distribution, then it implies that the dispersion is much less. If, on the

contrary, the Lorenz curve is farther away from the line of equal distribution, it implies that the dispersion is considerable.

The Lorenz curve is a simple graphical device to show the disparities of distribution in any phenomenon. It is, used in business and economics to represent inequalities in income, wealth, production, savings, and so on.

Figure 3.1 shows two Lorenz curves by way of illustration. The straight line AB is a line of equal distribution, whereas AEB shows complete inequality. Curve ACB and curve ADB are the Lorenz curves.

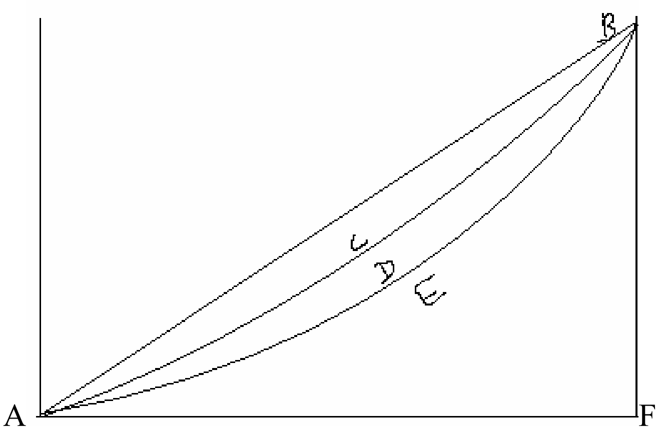


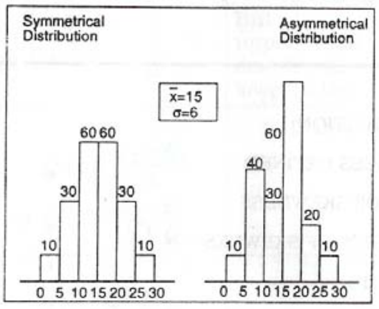
Figure 3.1: Lorenz Curve

As curve ACB is nearer to the line of equal distribution, it has more equitable distribution of income than curve ADB. Assuming that these two curves are for the same company, this may be interpreted in a different manner. Prior to taxation, the curve ADB showed greater inequality in the income of its employees. After the taxation, the company’s data resulted into ACB curve, which is closer to the line of equal distribution. In other words, as a result of taxation, the inequality has reduced.

3.10 SKEWNESS: MEANING AND DEFINITIONS

In the above paragraphs, we have discussed frequency distributions in detail. It may be repeated here that frequency distributions differ in three ways: Average value, Variability or dispersion, and Shape. Since the first two, that is, average value and variability or dispersion have already been discussed in previous chapters, here our main spotlight will be on the shape of frequency distribution. Generally, there are two comparable characteristics called skewness and kurtosis that help us to understand a distribution. Two distributions may have the same mean and standard deviation but may differ widely in their overall appearance as can be seen from the following:

In both these distributions the value of mean and standard deviation is the same ($\overline{X}$ = 15, σ = 5). But it does not imply that the distributions are alike in nature. The distribution on the left-hand side is a symmetrical one whereas the distribution on the right-hand side is symmetrical or skewed. Measures of skewness help us to distinguish between different types of distributions.



Some important definitions of skewness are as follows:

1. "When a series is not symmetrical it is said to be asymmetrical or skewed."
   -Croxton & Cowden.

2. "Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution." -Morris Hamburg.

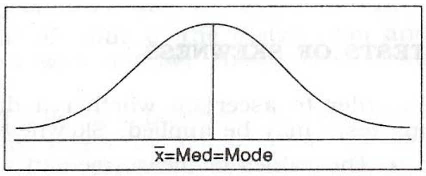
3. "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness."
   -Simpson & Kalka

4. "A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other-to left or right." -Garrett

The above definitions show that the term 'skewness' refers to lack of symmetry" i.e., when a distribution is not symmetrical (or is asymmetrical) it is called a skewed distribution.

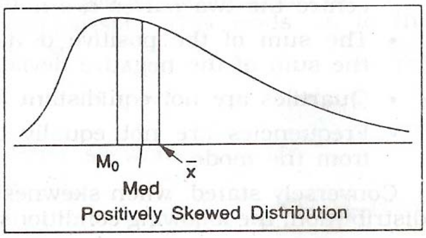
The concept of skewness will be clear from the following three diagrams showing a symmetrical distribution. a positively skewed distribution and a negatively skewed distribution.

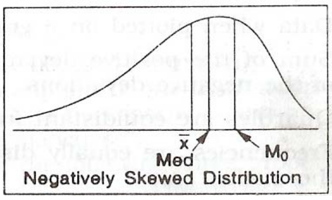
1. **Symmetrical Distribution.** It is clear from the diagram (a) that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the centre point of the curve.



2. **Asymmetrical Distribution.** A distribution, which is not symmetrical, is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed as would be clear from the diagrams (b) and (c).

3. **Positively Skewed Distribution.** In the positively skewed distribution the value of the mean is maximum and that of mode least-the median lies in between the two as is clear from the diagram (b).





4. **Negatively Skewed Distribution.** The following is the shape of negatively skewed distribution. In a negatively skewed distribution the value of mode is maximum and that of mean 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 high-value end of the curve (the right-hand side) than they are on the low-value end. In the negatively skewed distribution the position is reversed, i.e. the excess tail is on the left-hand side. It should be noted that in moderately symmetrical distributions the interval between the mean and the median is approximately one-third of the interval between the mean and the mode. It is this relationship, which provides a means of measuring the degree of skewness.

3.11 TESTS OF SKEWNESS

In order to ascertain whether a distribution is skewed or not the following tests may be applied. Skewness is present if:

1. The values of mean, median and mode do not coincide.

2. When the data are plotted on a graph they do not give the normal bell-shaped form i.e. when cut along a vertical line through the centre the two halves are not equal.

3. The sum of the positive deviations from the median is not equal to the sum of the negative deviations.

4. Quartiles are not equidistant from the median.

5. Frequencies are not equally distributed at points of equal deviation from the mode.

On the contrary, when skewness is absent, i.e. in case of a symmetrical distribution, the following conditions are satisfied:

1. The values of mean, median and mode coincide.

2. Data when plotted on a graph give the normal bell-shaped form.

3. Sum of the positive deviations from the median is equal to the sum of the negative deviations.

4. Quartiles are equidistant from the median.

5. Frequencies are equally distributed at points of equal deviations from the mode.

3.12 MEASURES OF SKEWNESS

There are four measures of skewness, each divided into absolute and relative measures. The relative measure is known as the coefficient of skewness and is more frequently used than the absolute measure of skewness. Further, when a comparison between two or more distributions is involved, it is the relative measure of skewness, which is used. The measures of skewness are: (i) Karl Pearson's measure, (ii) Bowley's measure, (iii) Kelly's measure, and (iv) Moment's measure. These measures are discussed briefly below:

3.12.1 KARL PEARON'S MEASURE

The formula for measuring skewness as given by Karl Pearson is as follows:

Skewness = Mean - Mode

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

In case the mode is indeterminate, the coefficient of skewness is:

$$Sk_p = \frac{\text{Mean - (3 Median - 2 Mean)}}{\textbf{Standard deviation}}$$

$$Sk_p = \frac{\text{3(Mean - Median)}}{\text{Standard deviation}}$$

Now this formula is equal to the earlier one.

$$\frac{\text{3(Mean - Median)}}{\text{Standard deviation}} \quad \frac{\text{Mean - Mode}}{\text{Standard deviation}}$$

Or 3 Mean - 3 Median = Mean - Mode

Or Mode = Mean - 3 Mean + 3 Median

Or Mode = 3 Median - 2 Mean

The direction of skewness is determined by ascertaining whether the mean is greater than the mode or less than the mode. If it is greater than the mode, then skewness is positive. But when the mean is less than the mode, it is negative. The difference between the mean and mode indicates the extent of departure from symmetry. It is measured in standard deviation units, which provide a measure independent of the unit of measurement. It may be recalled that this observation was made in the preceding chapter while discussing standard deviation. The value of coefficient of skewness is zero, when the distribution is symmetrical. Normally, this coefficient of skewness lies between $\underline{+}1$. If the mean is greater than the mode, then the coefficient of skewness will be positive, otherwise negative.

Example 3.11: Given the following data, calculate the Karl Pearson's coefficient of skewness: $\sum x = 452$ $\sum x^2 = 24270$ Mode = 43.7 and N = 10

Solution:

Pearson's coefficient of skewness is:

$$Sk_P = \frac{\text{Mean - Mode}}{\textbf{Standard deviation}}$$

Mean $(\bar{x})$= $\frac{\sum X}{N} = \frac{452}{10} = 45.2$

SD $(\sigma)=\sqrt{\frac{\sum x^2}{N}-\left(\frac{\sum x}{N}\right)^2}$ $(\sigma)=\sqrt{\frac{\sum x^2}{N}-\left(\frac{\sum x}{N}\right)^2}$

$(\sigma)=\sqrt{\frac{24270}{10}-\left(\frac{452}{10}\right)^2}$ $=\sqrt{2427-(45.2)^2}$ $=19.59$

Applying the values of mean, mode and standard deviation in the above formula,

$Sk_p = \frac{45.2-43.7}{19.59}$

=0.08

This shows that there is a positive skewness though the extent of skewness is marginal.

Example 3.12: From the following data, calculate the measure of skewness using the mean, median and standard deviation:

X	10 - 20	20 - 30	30 - 40	40 - 50	50-60	60 - 70	70 - 80
f	18	30	40	55	38	20	16

Solution:

x	*MVx*	d_x	*f*	fd_x	fd_X^2	*cf*
10 - 20	15	-3	18	-54	162	18
20 - 30	25	-2	30	-60	120	48
30 - 40	35	-1	40	-40	40	88
40-50	45=a	0	55	0	0	143
50 - 60	55	1	38	38	38	181
60 - 70	65	2	20	40	80	201
70 - 80	75	3	16	48	144	217
		Total	**217**	**-28**	**584**	

a = Assumed mean = 45, *cf* = Cumulative frequency, *dx* = Deviation from assumed mean, and *i* = 10

$$\bar{x} = a + \frac{\sum fdx}{N} \times i$$

$$= 45 - \frac{28}{217} \times 10 = 43.71$$

Median= $l_1 + \frac{l_2 - l_1}{f_1}(m - c)$

Where m = $(N + 1)/2^{\text{th}}$ item

= $(217 + 1)/2 = 109^{\text{th}}$ item

Median $= 40 - \frac{50 - 40}{55}(109 - 88)$

$$= 40 + \frac{10}{55} \times 21$$

= 43.82

SD = $\sqrt{\frac{\sum fd_x^2}{\sum f} - \left(\frac{\sum fd_x}{\sum f}\right)^2} \times 10 = \sqrt{\frac{584}{217} - \left(\frac{-28}{217}\right)^2} \times 10$

= $\sqrt{2.69 - 0.016} \times 10 = 16.4$

Skewness = 3 (Mean - Median)

= 3 (43.71 - 43.82)

= 3 x -0.011

= -0.33

Coefficient of skewness

$\frac{\text{Skewness}}{\text{SD}}$ or

$= \frac{-0.33}{16.4}$

$= -0.02$

The result shows that the distribution is negatively skewed, but the extent of skewness is extremely negligible.

3.12.2 Bowley's Measure

Bowley developed a measure of skewness, which is based on quartile values. The formula for measuring skewness is:

Skewness = $\frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$

Where Q_3 and Q_1 are upper and lower quartiles and M is the median. The value of this skewness varies between ± 1. In the case of open-ended distribution as well as where extreme values are found in the series, this measure is particularly useful. In a symmetrical distribution, skewness is zero. This means that Q_3 and Q_1 are positioned equidistantly from Q_2 that is, the median. In symbols, $Q_3 - Q_2 = Q_2 - Q_1'$ In contrast, when the distribution is skewed, then $Q_3 - Q_2$ will be different from $Q_2 - Q_1'$ When $Q_3 - Q_2$ exceeds $Q_2 - Q_1'$ then skewness is positive. As against this; when $Q_3 - Q_2$ is less than $Q_2 - Q_1'$ then skewness is negative. Bowley's measure of skewness can- be written as:

Skewness = $(Q_3 - Q_2) - (Q_2 - Q_1$ or $Q_3 - Q_2 - Q_2 + Q_1$

Or $Q_3 + Q_1 - 2Q_2$ $(2Q_2$ is $2M)$

However, this is an absolute measure of skewness. As such, it cannot be used while comparing two distributions where the units of measurement are different. In view of this limitation, Bowley suggested a relative measure of skewness as given below:

$$\text{Relative Skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

$$= \frac{Q_3 - Q_2 - Q_2 - Q_1}{Q_3 - Q_2 + Q_2 - Q_1}$$

$$= \frac{Q_3 - Q_1 - 2Q_2}{Q_3 - Q_1}$$

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

Example 3.13: For a distribution, Bowley's coefficient of skewness is - 0.56, Q_1*=16.4* and Median=24.2. What is the coefficient of quartile deviation?

Solution:

Bowley's coefficient of skewness is: $\text{Sk}_\text{B} = \dfrac{Q_3 - Q_1 - 2M}{Q_3 - Q_1}$

Substituting the values in the above formula,

$$\text{Sk}_\text{B} = \frac{\text{Q}_3 + 16.4 - (2 \text{ x } 24.2)}{Q_3 - 16.4}$$

$$-0.56 = \frac{\text{Q}_3 + 16.4 - 48.4}{Q_3 - 16.4}$$

or $- 0.56\, (Q_3\text{-}16.4) = Q_3\text{-}32$

or $- 0.56\, Q_3 + 9.184 = Q_3\text{-}32$

or $- 0.56\, Q_3 - Q_3 = -32 - 9.184$

$- 1.56\, Q_3 = - 41.184$

$$Q3 = \frac{-41.184}{1.56} = 26.4$$

Now, we have the values of both the upper and the lower quartiles.

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{26.4 - 16.4}{26.4 + 16.4} = \frac{10}{42.8} = 0.234 \text{ Approx.}$$

Example 3.14: Calculate an appropriate measure of skewness from the following data:

Value in Rs	Frequency
Less than 50	40
50 - 100	80
100 - 150	130
150 – 200	60
200 and above	30

Solution: It should be noted that the series given in the question is an open-ended series. As such, Bowley's coefficient of skewness, which is based on quartiles, would be the most appropriate measure of skewness in this case. In order to calculate the quartiles and the median, we have to use the cumulative frequency. The table is reproduced below with the cumulative frequency.

Value in Rs	Frequency	Cumulative Frequency
Less than 50	40	40
50 - 100	80	120
100 - 150	130	250
150 - 200	60	310
200 and above	30	340

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

Now $m = (\frac{n+1}{4})$ item $= \frac{341}{4}$ = 85.25, which lies in 50 - 100 class

$$Q_1 = 50 + \frac{100-50}{80}(85.25 - 40) = 78.28$$

$M = (\frac{n+1}{4})$ item $= \frac{341}{4}$ = 170.25, which lies in 100 - 150 class

M= $100 + \frac{150-100}{130}(170.5-120) = 119.4$

$Q_3 = l_1 + \frac{l_2 - l_1}{f_1}(m-c)$

$m = 3(341) \div 4 = 255.75$

$Q_3 = 150 + \frac{200-150}{60}(255.75-250) = 154.79$

Bowley's coefficient of skewness is:

$$\frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{154.79 + 78.28 - (2 \text{ x } 119.4)}{154.79 - 78.28} = \frac{-5.73}{76.51}$$

= - 0.075 approx.

This shows that there is a negative skewness, which has a very negligible magnitude.

3.12.3 Kelly's Measure

Kelly developed another measure of skewness, which is based on percentiles. The formula for measuring skewness is as follows:

Coefficient of skewness = $\frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$

Or, $\frac{D_1 + D_9 - 2M}{D_9 - D_1}$

Where *P* and *D* stand for percentile and decile respectively. In order to calculate the coefficient of skewness by this formula, we have to ascertain the values of 10th, 50th and 90th percentiles. Somehow, this measure of skewness is seldom used. All the same, we give an example to show how it can be calculated.

Example 3.15: Use Kelly's measure to calculate skewness.

Class Intervals	*f*	*cf*
10 - 20	18	18
20 - 30	30	48

30- 40	40	88
40- 50	55	143
50 - 60	38	181
60 – 70	20	201
70 - 80.	16	217

Solution: Now we have to calculate *P_{10} P_{30}* and *P_{90}.*

$P_{IO} = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$, where m = *(n* + 1)/10th item

$\frac{217+1}{10} = 21.8th$ item

This lies in the 20 - 30 class.

$$20 + \frac{30-20}{30}(21.8-18) = 20 + \frac{10 \times 3.8}{30} = 21.27 approx.$$

P_{50} (median): where m = *(n* + 1)/2th item = $\frac{217+1}{2}$ = 109th item

This lies in the class 40 - 50. Applying the above formula:

$$40 + \frac{50-40}{55}(109-88) = 40 + \frac{10 \times 21}{55} \times 21 = 43.82 approx.$$

$P_{90:}$ here m = 90 (217 + 1)/100th item = 196.2th item

This lies in the class 60 - 70. Applying the above formula:

$$60 + \frac{70-60}{20}(196.2-181) = 60 + \frac{10 \times 15.2}{20} = 67.6 approx.$$

Kelley's skewness

$$Sk_K \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

$$= \frac{67.6 - (2 \text{ x } 43.82) + 21.27}{67.6 - 21.27}$$

$$= \frac{88.87 - 87.64}{46.63}$$

= 0.027

This shows that the series is positively skewed though the extent of skewness is extremely negligible. It may be recalled that if there is a perfectly symmetrical distribution, then the skewness will be zero. One can see that the above answer is very close to zero.

3.13 MOMENTS

In mechanics, the term *moment* is used to denote the rotating effect of a force. In Statistics, it is used to indicate peculiarities of a frequency distribution. The utility of moments lies in the sense that they indicate different aspects of a given distribution. Thus, by using moments, we can measure the central tendency of a series, dispersion or variability, skewness and the peakedness of the curve. The moments about the actual arithmetic mean are denoted by μ. The first four moments about mean or *central moments* are as follows:

First moment $\mu_1 = \frac{1}{N}\sum\left(x_1 - \bar{x}\right)$

Second moment $\mu_2 = \frac{1}{N}\sum\left(x_1 - \bar{x}\right)^2$

Third moment $\mu_3 = \frac{1}{N}\sum\left(x_1 - \bar{x}\right)^3$

Fourth moment $\mu_3 = \frac{1}{N}\sum\left(x_1 - \bar{x}\right)^4$

These moments are in relation to individual items. In the case of a frequency distribution, the first four moments will be:

First moment $\mu_1 = \frac{1}{N}\sum fi\left(x_1 - \bar{x}\right)$

Second moment $\mu_2 = \frac{1}{N}\sum fi\left(x_1 - \bar{x}\right)^2$

Third moment $\mu_3 = \frac{1}{N}\sum fi\left(x_1 - \bar{x}\right)^3$

Fourth moment $\mu_3 \quad = \quad \frac{1}{N}\sum fi\left(x_1-\bar{x}\right)^4$

It may be noted that the first central moment is zero, that is, $\mu = 0$.

The second central moment is $\mu_2=\sigma$, indicating the variance.

The third central moment μ_3 is used to measure skewness. The fourth central moment gives an idea about the Kurtosis.

Karl Pearson suggested another measure of skewness, which is based on the third and second central moments as given below:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

Example 3.16: Find the (a) first, (b) second, (c) third and (d) fourth moments for the set of numbers 2,3,4,5 and 6.

Solution:

(a) $$\bar{x} = \frac{\sum x}{N} = \frac{2+3+4+5+6}{5} = \frac{20}{5} = 4$$

(b) $$\bar{x} = \frac{\sum x^2}{N} = \frac{2^2+3^2+4^2+5^2+6^2}{5}$$

$$= \frac{4+9+16+25+36}{5} = 18$$

(c) $$\bar{x} = \frac{\sum x^3}{N} = \frac{2^3+3^3+4^3+5^3+6^3}{5}$$

$$= \frac{8+27+64+125+216}{5} = 88$$

(d) $$\bar{x} = \frac{\sum x^4}{N} = \frac{2^4+3^4+4^4+5^4+6^4}{5}$$

$$= \frac{16+81+256+625+1296}{5} = 454.8$$

Example 3.17: Using the same set of five figures as given in Example 3.7, find the (a) first, (b) second, (c) third and (d) fourth moments about the mean.

Solution:

$$m_1 \quad = (x-\bar{x}) = \frac{\sum(x-\bar{x})}{N} = \frac{(2-4)+(3-4)+(4-4)+(5-4)+(6-4)}{5}$$

$$= \frac{-2-1+0+1+2}{5} = 0$$

$$m_2 = (x-\bar{x})^2 = \frac{\sum(x-\bar{x})^2}{N} = \frac{(2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2}{5}$$

$$= \frac{(-2)^2+(\_1)^2+0^2+1^2+2^2}{5}$$

$$= \frac{4+1+0+1+4}{5} = 2.$$ It may be noted that m_2 is the variance

$$m_3 = = (x-\bar{x})^3 = \frac{\sum(x-\bar{x})^3}{N} = \frac{(2-4)^3+(3-4)^3+(4-4)^3+(5-4)^3+(6-4)^3}{5}$$

$$= \frac{(-2)^3+(\_1)^3+0^3+1^3+2^3}{5} = \frac{-8-1+0+1+8}{5} = 0$$

$$m_4 = = (x-\bar{x})^4 = \frac{\sum(x-\bar{x})^4}{N} = \frac{(2-4)^4+(3-4)^4+(4-4)^4+(5-4)^4+(6-4)^4}{5}$$

$$= \frac{(-2)^4+(\_1)^4+0^4+1^4+2^4}{5}$$

$$= \frac{16+1+0+1+016}{5} = 6.8$$

Example 3.18: Calculate the first four central moments from the following data:

Class interval	50-60	60-70	70-80	80-90	90-100
Frequency	5	12	20	7	6

Solution:

Class Interval	f	MV	d from 75	d/10	fd	fd^2	fd^3	fd^4
50- 60	5	55	-20	-2	-10	20	-40	80

60- 70	12	65	-10	-1	-12	12	-12	12
70- 80	20	75	0	0	0	0	0	0
80- 90	7	85	10	1	7	7	7	7
90-100	6	95	20	2	12	24	48	96
Total	50				-3		-4	195

$$\mu_1' = \frac{\sum fd \times i}{N} = \frac{-3 \times 10}{50} = -0.6$$

$$\mu_2' = \frac{\sum fd^2 \times i}{N} = \frac{63 \times 10}{50} = 12.6$$

$$\mu_2' = \frac{\sum fd^3 \times i}{N} = \frac{-4 \times 10}{50} = 0.8$$

$$\mu_2' = \frac{\sum fd^4 \times i}{N} = \frac{195 \times 10}{50} = 19$$

Moments about Mean

$\mu_1 = \mu_1' - \mu_1' = -0.6-(-0.6) = 0$

$\mu_2 = \mu_2' - \mu_1'^2 = 10-(-0.6)^2 = 10-3.6 = 6.4$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -0.8-3(12.6)(-0.6)+2(-0.6)^3$

$= -0.8 + 22.68 + 0.432 = 22.312$

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

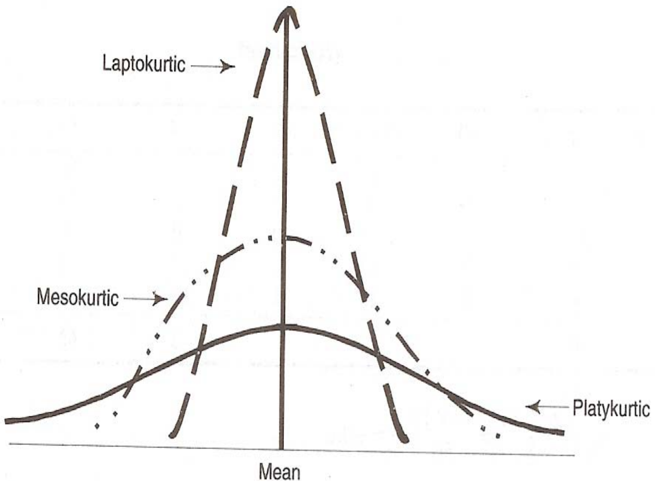
$= 19 + 4(-0.8)(-0.6) + 6(10)(-0.6)^2 - 3(-0.6)^4$

$= 19 + 1.92 + 21.60 - 0.3888$

$= 42.1312$

3.14 KURTOSIS

Kurtosis is another measure of the shape of a frequency curve. It is a Greek word, which means bulginess. While skewness signifies the extent of asymmetry, kurtosis measures the degree of peakedness of a frequency distribution. Karl Pearson classified curves into three types on the basis of the shape of their peaks. These are mesokurtic, leptokurtic and platykurtic. These three types of curves are shown in figure below:



It will be seen from Fig. 3.2 that mesokurtic curve is neither too much flattened nor too much peaked. In fact, this is the frequency curve of a normal distribution. Leptokurtic curve is a more peaked than the normal curve. In contrast, platykurtic is a relatively flat curve. The coefficient of kurtosis as given by Karl Pearson is $\beta_2 = \mu_4/\mu_2^2$. In case of a normal distribution, that is, mesokurtic curve, the value of $\beta_2=3$. If β_2 turn out to be > 3, the curve is called a leptokurtic curve and is more peaked than the normal curve. Again, when $\beta_2 < 3$, the curve is called a platykurtic curve and is less peaked than the normal curve. The measure of kurtosis is very helpful in the selection of an appropriate average. For example, for normal distribution, mean is most appropriate; for a leptokurtic distribution, median is most appropriate; and for platykurtic distribution, the quartile range is most appropriate.

Example 3.19: From the data given in Example 3.18, calculate the kurtosis.

Solution: For this, we have to calculate β_2 This can be done by using the formula $\beta_2 = \mu_4/\mu_2^2$. In the preceding example, values of μ_4 and μ_2 are given. Hence, $\beta_2 = 42.1312 \div (6.4)^2 = 1.03$.

As β_2. < 3, the distribution is platykurtic.

Another measure of kurtosis is based on both quartiles and percentiles and is given by the following formula:

$$K = \frac{Q}{P_{90} - P_{10}}$$

Where K = kurtosis, $Q = ½ (Q_3 – Q_1)$ is the semi-interquartile range; P_{90} is 90th percentile and P_{10} is the 10th percentile. This is also known as the percentile *coefficient of kurtosis.* In case of the normal distribution, the value of K is 0.263.

Example 3.20: From the data given below, calculate the percentile coefficient of kurtosis.

Daily Wages in Rs.	**Number of Workers**	**cf**
50- 60	10	10
60-70	14	24
70-80	18	42
80 - 90	24	66
90-100	16	82
100 -110	12	94
110 - 120	6	100
Total	100	

Solution: It may be noted that the question involved first two columns and in order to calculate quartiles and percentiles, cumulative frequencies have been shown in column three of the above table.

$$Q_1 \quad = \quad l_1 \frac{l_2 - l_1}{f_1}(m - c)$$, where $m = (n + 1)/4$th item, which is = 25.25th item

This falls in 70 - 80 class interval.

$$= \quad 70 + \frac{80 - 70}{18}(25.25 - 24) = 70.69$$

$$Q3 \quad = \quad l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$, where $m = 75.75$

This falls in 90 - 100 class interval.

$$= \quad 90 + \frac{100 - 90}{16} (75.75 - 66) = 96.09$$

P$_{I0}$ = $l_1 + \frac{l_2 - l_1}{f_1}(m-c)$, where m = 10.1

This falls in 60 - 70 class interval.

= $60 + \frac{70-60}{14}$ (10.01 -10) = 60.07

P$_{90}$ = $l_1 + \frac{l_2 - l_1}{f_1}(m-c)$, where m = 90.9

This falls in 100 - 110 class interval.

= $100 + \frac{110-100}{12}$ (90.9 - 82) = 107.41

$$K = \frac{Q}{P_{90} - P_{10}}$$

$$= \frac{1/2(Q_3 - Q_1)}{P_{90} - P_{10}}$$

$$= \frac{\frac{1}{2}\ (96.09 - 70.69)}{107.41 - 60.07}$$

= 0.268

It will be seen that the above distribution is very close to normal distribution as the value of K is 0.268, which is extremely close to 0.263.

Objectives : *The overall objective of this lesson is to give you an understanding of bivariate linear correlation, there by enabling you to understand the importance as well as the limitations of correlation analysis.*

Structure

4.1 Introduction

4.2 What is Correlation?

4.3 Correlation Analysis

- 4.3.1 Scatter Diagram
- 4.3.2 Correlation Graph
- 4.3.3 Pearson's Coefficient of Correlation
- 4.3.4 Spearman's Rank Correlation
- 4.3.5 Concurrent Deviation Method

4.4 Limitations of Correlation Analysis

...if we have information on more than one variables, we might be interested in seeing if there is any connection - any association - between them.

4.1 INTRODUCTION

Statistical methods of measures of central tendency, dispersion, skewness and kurtosis are helpful for the purpose of comparison and analysis of distributions involving only one variable *i.e.* univariate distributions. However, describing the relationship between two or more variables, is another important part of statistics.

In many business research situations, the key to decision making lies in understanding the relationships between two or more variables. *For example*, in an effort to predict the behavior of the bond market, a broker might find it useful to know whether the interest rate of bonds is related to the prime interest rate. While studying the effect of advertising on sales, an account executive may find it useful to know whether there is a strong relationship between advertising dollars and sales dollars for a company.

The statistical methods of ***Correlation*** (discussed in the present lesson) and ***Regression*** (to be discussed in the next lesson) are helpful in knowing the relationship between two or more variables which may be related in same way, *like* interest rate of bonds and prime interest rate; advertising expenditure and sales; income and consumption; crop-yield and fertilizer used; height and weights and so on.

In all these cases involving two or more variables, we may be interested in seeing:

- if there is any association between the variables;
- if there is an association, is it strong enough to be useful;
- if so, what form the relationship between the two variables takes;
- how we can make use of that relationship for predictive purposes, that is, forecasting; and
- how good such predictions will be.

Since these issues are inter related, correlation and regression analysis, as two sides of a single process, consists of methods of examining the relationship between two or more variables. If two (or more) variables are correlated, we can use information about one (or more) variable(s) to predict the value of the other variable(s), and can measure the error of estimations - *a job of regression analysis.*

4.2 WHAT IS CORRELATION?

Correlation is a measure of association between two or more variables. When two or more variables very in sympathy so that movement in one tends to be accompanied by corresponding movements in the other variable(s), they are said to be correlated.

> *"The correlation between variables is a measure of the nature and degree of association between the variables".*

As *a measure of the degree of relatedness of two variables,* correlation is widely used in exploratory research when the objective is to locate variables that might be related in some way to the variable of interest.

4.2.1 TYPES OF CORRELATION

Correlation can be classified in several ways. The important ways of classifying correlation are:

(i) Positive and negative,

(ii) Linear and non-linear (curvilinear) and

(iii) Simple, partial and multiple.

Positive and Negative Correlation

If both the variables move in the same direction, we say that there is a positive correlation, *i.e.,* if one variable increases, the other variable also increases on an average or if one variable decreases, the other variable also decreases on an average.

On the other hand, if the variables are varying in opposite direction, we say that it is a case of negative correlation; *e.g.,* movements of demand and supply.

Linear and Non-linear (Curvilinear) Correlation

If the change in one variable is accompanied by change in another variable in a constant ratio, it is a case of linear correlation. Observe the following data:

X	:	10	20	30	40	50
Y	:	25	50	75	100	125

The ratio of change in the above example is the same. It is, thus, a case of linear correlation. If we plot these variables on graph paper, all the points will fall on the same straight line.

On the other hand, if the amount of change in one variable does not follow a constant ratio with the change in another variable, it is a case of non-linear or curvilinear correlation. If a couple of figures in either series X or series Y are changed, it would give a non-linear correlation.

Simple, Partial and Multiple Correlation

The distinction amongst these three types of correlation depends upon the number of variables involved in a study. If only two variables are involved in a study, then the correlation is said to be simple correlation. When three or more variables are involved in a study, then it is a problem of either partial or multiple correlation. In multiple correlation, three or more variables are studied simultaneously. But in partial correlation we consider only two variables influencing each other while the effect of other variable(s) is held constant.

Suppose we have a problem comprising three variables X, Y and Z. X is the number of hours studied, Y is I.Q. and Z is the number of marks obtained in the examination. In a multiple correlation, we will study the relationship between the marks obtained (Z) and the two variables, number of hours studied (X) and I.Q. (Y). In contrast, when we study the relationship between *X* and *Z,* keeping an average I.Q. (*Y*) as constant, it is said to be a study involving partial correlation.

In this lesson, we will study linear correlation between two variables.

4.2.2 CORRELATION DOES NOT NECESSARILY MEAN CAUSATION

The correlation analysis, in discovering the nature and degree of relationship between variables, does not necessarily imply any cause and effect relationship between the variables. Two variables may be related to each other but this does not mean that one variable causes the other. *For example*, we may find that logical reasoning and creativity are correlated, but that does not mean if we could increase peoples' logical reasoning ability, we would produce greater creativity. We need to conduct an actual experiment to unequivocally demonstrate a causal relationship. But if it is true that influencing someones' logical reasoning ability does influence their creativity, then the two variables must be correlated with each other. **In other words, *causation always implies correlation, however converse is not true.***

Let us see some situations-

1. The correlation may be due to chance particularly when the data pertain to a small sample. A small sample bivariate series may show the relationship but such a relationship may not exist in the universe.
2. It is possible that both the variables are influenced by one or more other variables. For example, expenditure on food and entertainment for a given number of households show a positive relationship because both have increased over time. But, this is due to rise in family incomes over the same period. In other words, the two variables have been influenced by another variable - increase in family incomes.

3. There may be another situation where both the variables may be influencing each other so that we cannot say which is the cause and which is the effect. *For example,* take the case of price and demand. The rise in price of a commodity may lead to a decline in the demand for it. Here, price is the cause and the demand is the effect. In yet another situation, an increase in demand may lead to a rise in price. Here, the demand is the cause while price is the effect, which is just the reverse of the earlier situation. In such situations, it is difficult to identify which variable is causing the effect on which variable, as both are influencing each other.

The foregoing discussion clearly shows that correlation does not indicate any causation or functional relationship. ***Correlation coefficient is merely a mathematical relationship and this has nothing to do with cause and effect relation.*** It only reveals co-variation between two variables. Even when there is no cause-and-effect relationship in bivariate series and one interprets the relationship as causal, such a correlation is called ***spurious*** or ***non-sense correlation***. Obviously, this will be misleading. As such, one has to be very careful in correlation exercises and look into other relevant factors before concluding a cause-and-effect relationship.

4.3 CORRELATION ANALYSIS

Correlation Analysis is a statistical technique used to indicate the nature and degree of relationship existing between one variable and the other(s). It is also used along with regression analysis to measure how well the regression line explains the variations of the dependent variable with the independent variable.

The commonly used methods for studying linear relationship between two variables involve both graphic and algebraic methods. Some of the widely used methods include:

1. Scatter Diagram
2. Correlation Graph

3. Pearson’s Coefficient of Correlation

4. Spearman’s Rank Correlation

5. Concurrent Deviation Method

4.3.1 SCATTER DIAGRAM

This method is also known as Dotogram or Dot diagram. Scatter diagram is one of the simplest methods of diagrammatic representation of a bivariate distribution. Under this method, both the variables are plotted on the graph paper by putting dots. The diagram so obtained is called "Scatter Diagram". By studying diagram, we can have rough idea about the nature and degree of relationship between two variables. The term scatter refers to the spreading of dots on the graph. We should keep the following points in mind while interpreting correlation:

- if the plotted points are very close to each other, it indicates high degree of correlation. If the plotted points are away from each other, it indicates low degree of correlation.

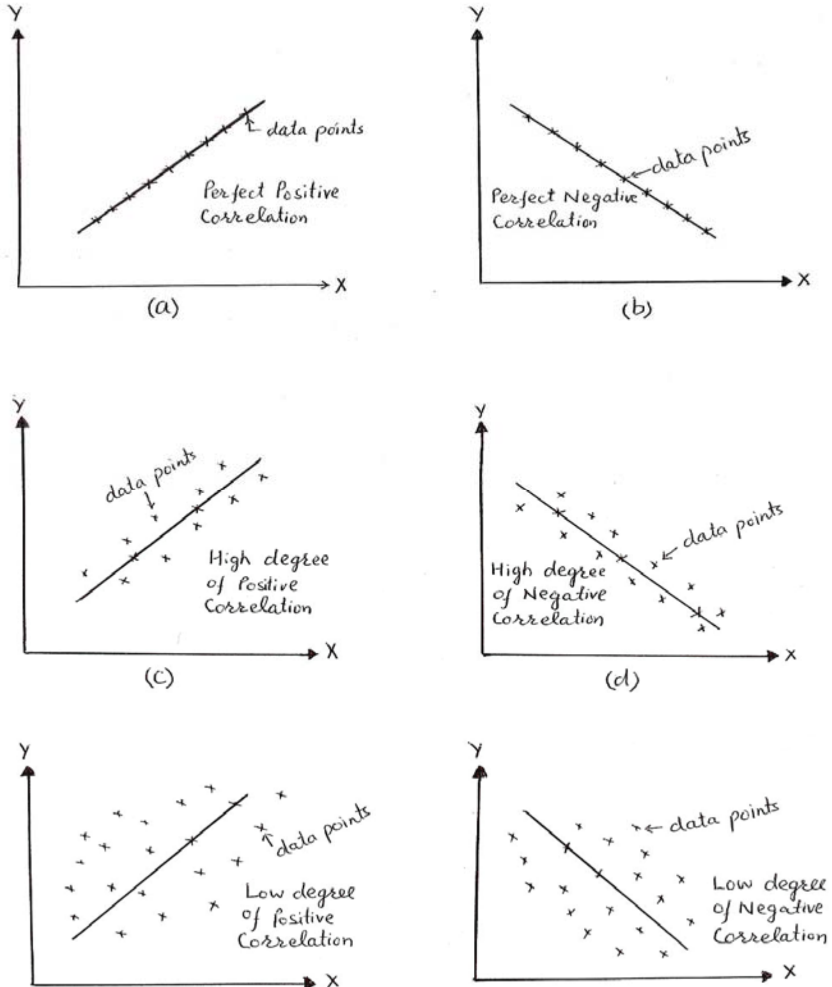


Figure 4-1 Scatter Diagrams

- if the points on the diagram reveal any trend (either upward or downward), the variables are said to be correlated and if no trend is revealed, the variables are uncorrelated.
- if there is an upward trend rising from lower left hand corner and going upward to the upper right hand corner, the correlation is positive since this reveals that the values of the two variables move in the same direction. If, on the other hand, the points depict a downward trend from the upper left hand corner to the lower right hand corner, the correlation is negative since in this case the values of the two variables move in the opposite directions.
- in particular, if all the points lie on a straight line starting from the left bottom and going up towards the right top, the correlation is perfect and positive, and if all the points like on a straight line starting from left top and coming down to right bottom, the correlation is perfect and negative.

The various diagrams of the scattered data in Figure 4-1 depict different forms of correlation.

Example 4-1

Given the following data on sales (in thousand units) and expenses (in thousand rupees) of a firm for 10 month:

Month :	J	F	M	A	M	J	J	A	S	O
Sales:	50	50	55	60	62	65	68	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

a) Make a Scatter Diagram

b) Do you think that there is a correlation between sales and expenses of the firm? Is it positive or negative? Is it high or low?

Solution:(a) The Scatter Diagram of the given data is shown in Figure 4-2

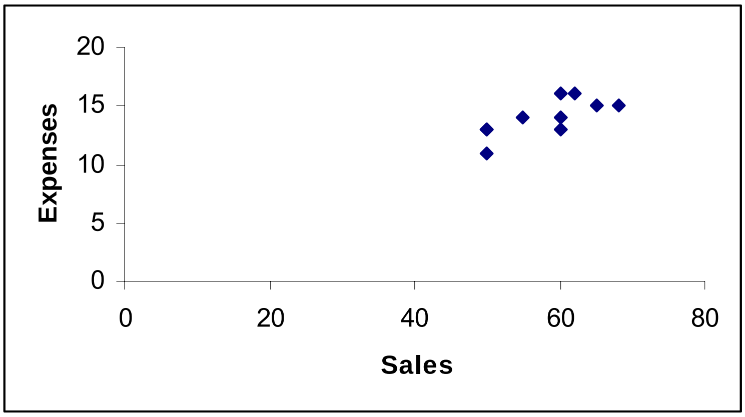


Figure 4.2 Scatter Diagram

(b) Figure 4-2 shows that the plotted points are close to each other and reveal an upward trend. So there is a high degree of positive correlation between sales and expenses of the firm.

4.3.2 CORRELATION GRAPH

This method, also known as Correlogram is very simple. The data pertaining to two series are plotted on a graph sheet. We can find out the correlation by examining the direction and closeness of two curves. If both the curves drawn on the graph are moving in the same direction, it is a case of positive correlation. On the other hand, if both the curves are moving in opposite direction, correlation is said to be negative. If the graph does not show any definite pattern on account of erratic fluctuations in the curves, then it shows an absence of correlation.

<u>Example 4-2</u>

Find out graphically, if there is any correlation between price yield per plot (qtls); denoted by *Y* and quantity of fertilizer used (kg); denote by *X*.

Plot No.:	1	2	3	4	5	6	7	8	9	10
Y:	3.5	4.3	5.2	5.8	6.4	7.3	7.2	7.5	7.8	8.3
X:	6	8	9	12	10	15	17	20	18	24

Solution: The Correlogram of the given data is shown in Figure 4-3

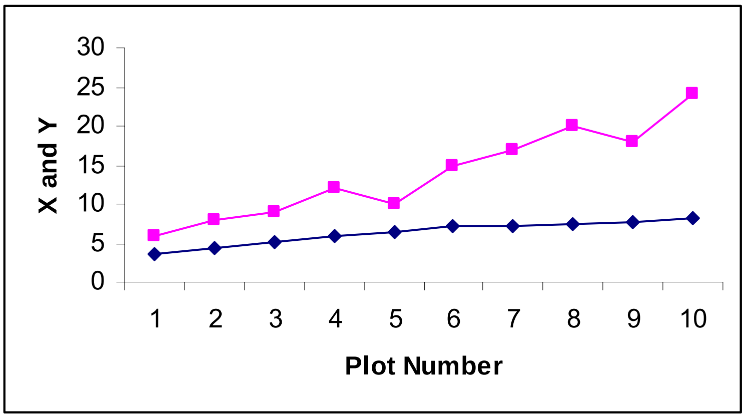


Figure 4-3 Correlation Graph

Figure 4-3 shows that the two curves move in the same direction and, moreover, they are very close to each other, suggesting a close relationship between price yield per plot (qtls) and quantity of fertilizer used (kg)

Remark: Both the Graphic methods - scatter diagram and correlation graph provide a *'feel for'* of the data – by providing visual representation of the association between the variables. These are readily comprehensible and enable us to form a fairly good, though rough idea of the nature and degree of the relationship between the two variables. However, these methods are unable to quantify the relationship between them. To quantify the extent of correlation, we make use of algebraic methods - which calculate correlation coefficient.

4.3.3 PEARSON'S COEFFICIENT OF CORRELATION

A mathematical method for measuring the intensity or the magnitude of *linear relationship* between two variables was suggested by Karl Pearson (1867-1936), a great British Biometrician and Statistician and, it is by far the most widely used method in practice.

Karl Pearson's measure, known as Pearsonian correlation coefficient between two variables *X* and *Y*, usually denoted by *r(X,Y)* or r_{xy} or simply *r* is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between *X* and *Y,* to the product of the standard deviations of *X* and *Y.*

Symbolically

$$r_{xy} = \frac{Cov(X,Y)}{S_x . S_y} \qquad \ldots\ldots\ldots\ldots(4.1)$$

when, $(X_1, Y_1); (X_2, Y_2); \ldots\ldots\ldots\ldots\ldots\ldots (X_n, Y_n)$ are *N* pairs of observations of the variables *X* and *Y* in a bivariate distribution,

$$Cov(X,Y) = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{N} \qquad \ldots\ldots\ldots\ldots(4.2a)$$

$$S_x = \sqrt{\frac{\sum (X - \overline{X})^2}{N}} \qquad \ldots\ldots\ldots\ldots(4.2b)$$

and $$S_y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{N}} \qquad \ldots\ldots\ldots\ldots(4.2c)$$

Thus by substituting *Eqs. (4.2)* in *Eq. (4.1),* we can write the Pearsonian correlation coefficient as

$$r_{xy} = \frac{\frac{1}{N}\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\frac{1}{N}\sum (X - \overline{X})^2}\sqrt{\frac{1}{N}\sum (Y - \overline{Y})^2}}$$

$$r_{xy} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2}\sqrt{\sum (Y - \overline{Y})^2}} \qquad \ldots\ldots\ldots\ldots(4.3)$$

If we denote, $d_x = X - \overline{X}$ and $d_y = Y - \overline{Y}$

Then $$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x{}^2 d_y{}^2}} \qquad \ldots\ldots\ldots\ldots(4.3a)$$

We can further simply the calculations of *Eqs. (4.2)*

We have

$$Cov(X,Y) = \frac{1}{N}\sum (X - \overline{X})(Y - \overline{Y})$$

$$= \frac{1}{N}\sum XY - \overline{X}\overline{Y}$$

$$= \frac{1}{N}\sum XY - \frac{\sum X}{N}\frac{\sum Y}{N}$$

$$= \frac{1}{N^2}\left[N\sum XY - \sum X \sum Y\right] \qquad \ldots\ldots\ldots\ldots(4.4)$$

and $$S_x^2 = \frac{1}{N}\sum (X - \overline{X})^2$$

$$= \frac{1}{N}\sum X^2 - (\overline{X})^2$$

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

$$= \frac{1}{N^2}\left[N\sum X^2 - \left(\sum X\right)^2\right] \qquad \ldots\ldots\ldots\ldots(4.5a)$$

Similarly, we have

$$S_y^2 = \frac{1}{N^2}\left[N\sum Y^2 - \left(\sum Y\right)^2\right] \qquad \ldots\ldots\ldots\ldots(4.5b)$$

So Pearsonian correlation coefficient may be found as

$$r_{xy} = \frac{\frac{1}{N^2}\left[N\sum XY - \sum X \sum Y\right]}{\sqrt{\frac{1}{N^2}\left[N\sum X^2 - \left(\sum X\right)^2\right]}\sqrt{\frac{1}{N^2}\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

or $$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}} \qquad \ldots\ldots\ldots\ldots(4.6)$$

Remark: *Eq. (4.3)* or *Eq. (4.3a)* is quite convenient to apply if the means $\overline{X}$ and $\overline{Y}$ come out to be integers. If $\overline{X}$ or/and $\overline{Y}$ is (are) fractional then the *Eq. (4.3)* or *Eq. (4.3a)* is quite cumbersome to apply, since the computations of $\sum(X-\overline{X})^2$, $\sum(Y-\overline{Y})^2$ and $\sum(X-\overline{X})(Y-\overline{Y})$ are quite time consuming and tedious. In such a case *Eq. (4.6)* may be used provided the values of *X* or/ and *Y* are small. But if *X* and *Y* assume large values, the calculation of $\sum X^2$,$\sum Y^2$ and $\sum XY$ is again quite time consuming.

Thus if *(i)* $\overline{X}$ and $\overline{Y}$ are fractional and *(ii) X* and *Y* assume large values, the *Eq. (4.3)* and *Eq. (4.6)* are not generally used for numerical problems. In such cases, the step deviation method where we take the deviations of the variables *X* and *Y* from any arbitrary points is used. We will discuss this method in the properties of correlation coefficient.

4.3.3.1 Properties of Pearsonian Correlation Coefficient

The following are important properties of Pearsonian correlation coefficient:

1. *Pearsonian correlation coefficient cannot exceed 1 numerically*. In other words it lies between –1 and +1. Symbolically,

$$-1 \leq r \leq 1$$

Remarks: *(i)* This property provides us a check on our calculations. If in any problem, the obtained value of *r* lies outside the limits ± 1, this implies that there is some mistake in our calculations.

(ii) The sign of *r* indicate the nature of the correlation. Positive value of *r* indicates positive correlation, whereas negative value indicates negative correlation. *r = 0* indicate absence of correlation.

(iii) The following table sums up the degrees of correlation corresponding to various values of *r:*

Value of *r*	**Degree of correlation**
±1	perfect correlation
±0.90 or more	very high degree of correlation
±0.75 to ±0.90	sufficiently high degree of correlation
±0.60 to ±0.75	moderate degree of correlation
±0.30 to ±0.60	only the possibility of a correlation
less than ±0.30	*possibly no correlation*
0	absence of correlation

2. *Pearsonian Correlation coefficient is independent of the change of origin and scale.* Mathematically, if given variables *X* and *Y* are transformed to new variables *U* and *V* by change of origin and scale, *i. e.*

$$U = \frac{X - A}{h} \quad \text{and} \quad V = \frac{Y - B}{k}$$

Where *A, B, h* and *k* are constants and *h > 0, k > 0;* then the correlation coefficient between *X* and *Y* is same as the correlation coefficient between *U* and *V i.e.,*

$$r(X,Y) = r(U, V) \Rightarrow r_{xy} = r_{uv}$$

Remark: This is one of the very important properties of the correlation coefficient and is extremely helpful in numerical computation of *r*. We had already stated that *Eq. (4.3)* and *Eq.(4.6)* become quite tedious to use in numerical problems if *X* and/or *Y* are in fractions or if *X* and *Y* are large. In such cases we can conveniently change the origin and scale (if possible) in *X* or/and *Y* to get new variables *U* and *V* and compute the correlation between *U* and *V* by the *Eq. (4.7)*

$$r_{xy} = r_{uv} = \frac{N\sum UV - \sum U \sum V}{\sqrt{N\sum U^2 - \left(\sum U\right)^2}\sqrt{N\sum V^2 - \left(V\right)^2}} \qquad \ldots\ldots\ldots\ldots(4.7)$$

3. *Two independent variables are uncorrelated but the converse is not true*

If X and Y are independent variables then

$$r_{xy} = 0$$

However, the converse of the theorem is not true *i.e.,* uncorrelated variables need not necessarily be independent. As an illustration consider the following bivariate distribution.

X	:	1	2	3	-3	-2	-1
Y	:	1	4	9	9	4	1

For this distribution, value of r will be 0.

Hence in the above example the variable X and Y are uncorrelated. But if we examine the data carefully we find that X and Y are not independent but are connected by the relation $Y = X^2$. The above example illustrates that uncorrelated variables need not be independent.

Remarks: One should not be confused with the words uncorrelation and independence. $r_{xy} = 0$ *i.e.,* uncorrelation between the variables X and Y simply implies the absence of any linear (straight line) relationship between them. They may, however, be related in some other form other than straight line *e.g.,* quadratic (as we have seen in the above example), logarithmic or trigonometric form.

4. *Pearsonian coefficient of correlation is the geometric mean of the two regression coefficients, i.e.*

$$r_{xy} = \pm\sqrt{b_{xy}.b_{yx}}$$

The signs of both the regression coefficients are the same, and so the value of r will also have the same sign.

This property will be dealt with in detail in the next lesson on Regression Analysis.

5. *The square of Pearsonian correlation coefficient is known as the coefficient of determination.*

   Coefficient of determination, which measures the percentage variation in the dependent variable that is accounted for by the independent variable, is a much better and useful measure for interpreting the value of *r*. This property will also be dealt with in detail in the next lesson.

4.3.3.2 Probable Error of Correlation Coefficient

The correlation coefficient establishes the relationship of the two variables. After ascertaining this level of relationship, we may be interested to find the extent upto which this coefficient is dependable. Probable error of the correlation coefficient is such a measure of testing the reliability of the observed value of the correlation coefficient, when we consider it as satisfying the conditions of the random sampling.

If *r* is the observed value of the correlation coefficient in a sample of *N* pairs of observations for the two variables under consideration, then the Probable Error, denoted by *PE* (*r*) is expressed as

$$PE(r) = 0.6745 \ \ SE(r)$$

or $$PE(r) = 0.6745 \frac{1-r^2}{\sqrt{N}}$$

There are two main functions of probable error:

1. ***Determination of limits***: The limits of population correlation coefficient are *r* ± *PE(r),* implying that if we take another random sample of the size *N* from the same population, then the observed value of the correlation coefficient in the second sample can be expected to lie within the limits given above, with 0.5 probability. When sample size *N* is small, the concept or value of *PE* may lead to wrong

conclusions. Hence to use the concept of *PE* effectively, sample size *N* it should be fairly large.

2. ***Interpretation of 'r':*** The interpretation of *'r'* based on *PE* is as under:

   - If $r < PE(r)$, there is no evidence of correlation, *i.e.* a case of insignificant correlation.
   - If $r > 6\ PE(r)$, correlation is significant. *If* $r < 6\ PE(r)$, it is insignificant.
   - If the probable error is small, correlation exist where $r > 0.5$

Example 4-3

Find the Pearsonian correlation coefficient between sales (in thousand units) and expenses (in thousand rupees) of the following 10 firms:

Firm:	1	2	3	4	5	6	7	8	9	10
Sales:	50	50	55	60	65	65	65	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

Solution: Let sales of a firm be denoted by *X* and expenses be denoted by *Y*

Calculations for Coefficient of Correlation

{Using *Eq. (4.3)* or *(4.3a)*}

Firm	X	Y	$d_x = X - \bar{X}$	$d_y = Y - \bar{Y}$	d_x^2	d_y^2	$d_x.d_y$
1	*50*	*11*	*-8*	*-3*	*64*	*9*	*24*
2	*50*	*13*	*-8*	*-1*	*64*	*1*	*8*
3	*55*	*14*	*-3*	*0*	*9*	*0*	*0*
4	*60*	*16*	*2*	*2*	*4*	*4*	*4*
5	*65*	*16*	*7*	*2*	*49*	*4*	*14*
6	*65*	*15*	*7*	*1*	*49*	*1*	*7*
7	*65*	*15*	*7*	*1*	*49*	*1*	*7*
8	*60*	*14*	*2*	*0*	*4*	*0*	*0*
9	*60*	*13*	*2*	*-1*	*4*	*1*	*-2*
10	*50*	*13*	*-8*	*-1*	*64*	*1*	*8*
	$\sum X$	$\sum Y$			$\sum d_x^2$	$\sum d_y^2$	$\sum d_x d_y$

	= 580	= 140		=360	=22	=70

$\overline{X} = \frac{\sum X}{N} = \frac{580}{10} = 58$ and $\overline{Y} = \frac{\sum Y}{N} = \frac{140}{10} = 14$

Applying the *Eq. (4.3a), w*e have, Pearsonian coefficient of correlation

$$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x{}^2 d_y{}^2}}$$

$$r_{xy} = \frac{70}{\sqrt{360x22}}$$

$$r_{xy} = \frac{70}{\sqrt{7920}}$$

$$r_{xy} = 0.78$$

The value of $r_{xy} = 0.78$*, indicate a high degree of positive correlation between sales and expenses.*

Example 4-4

The data on price and quantity purchased relating to a commodity for 5 months is given below:

Month :	January	February	March	April	May
Prices(Rs):	10	10	11	12	12
Quantity(Kg):	5	6	4	3	3

Find the Pearsonian correlation coefficient between prices and quantity and comment on its sign and magnitude.

Solution: Let price of the commodity be denoted by X and quantity be denoted by Y

Calculations for Coefficient of Correlation

{Using *Eq. (4.6)*}

Month	X	Y	X^2	Y^2	XY
1	*10*	*5*	*100*	*25*	*50*
2	*10*	*6*	*100*	*36*	*60*
3	*11*	*4*	*121*	*16*	*44*
4	*12*	*3*	*144*	*9*	*36*
5	*12*	*3*	*144*	*9*	*36*

	$\sum X$ =55	$\sum Y$ =21	$\sum X^2 = 609$	$\sum Y^2 = 95$	$\sum XY = 226$

Applying the *Eq. (4.6), w*e have, Pearsonian coefficient of correlation

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r_{xy} = \frac{5x226 - 55x21}{\sqrt{(5x609 - 55x55)(5x95 - 21x21)}}$$

$$r_{xy} = \frac{1130 - 1155}{\sqrt{20x34}}$$

$$r_{xy} = \frac{-25}{\sqrt{680}}$$

$$r_{xy} = -0.98$$

The negative sign of *r* indicate negative correlation and its large magnitude indicate a very high degree of correlation. So there is a high degree of negative correlation between prices and quantity demanded.

Example 4-5

Find the Pearsonian correlation coefficient from the following series of marks obtained by 10 students in a class test in mathematics (*X)* and in Statistics (*Y)*:

X:	45	70	65	30	90	40	50	75	85	60
Y:	35	90	70	40	95	40	60	80	80	50

Also calculate the Probable Error.

Solution:

Calculations for Coefficient of Correlation

{Using *Eq. (4.7)*}

X	Y	U	V	U^2	V^2	UV
45	*35*	*-3*	*-6*	*9*	*36*	*18*
70	*90*	*2*	*5*	*4*	*25*	*10*
65	*70*	*1*	*1*	*1*	*1*	*1*

30	*40*	*-6*	*-5*	*36*	*25*	*30*
90	*95*	*6*	*6*	*36*	*36*	*36*
40	*40*	*-4*	*-5*	*16*	*25*	*20*
50	*60*	*-2*	*-1*	*4*	*1*	*2*
75	*80*	*3*	*3*	*9*	*9*	*9*
85	*80*	*5*	*3*	*25*	*9*	*15*
60	*50*	*0*	*-3*	*0*	*9*	*0*
		$\sum U = 2$	$\sum V = -2$	$\sum U^2 = 140$	$\sum V^2 = 176$	$\sum UV \ = 141$

We have, defined variables *U* and *V* as

$$U = \frac{X - 60}{5} \qquad \text{and} \qquad V = \frac{Y - 65}{5}$$

Applying the *Eq. (4.7)*

$$r_{xy} = r_{uv} = \frac{N\sum UV - \left(\sum U \sum V\right)}{\sqrt{N\sum U^2 - \left(\sum U\right)^2}\sqrt{N\sum V^2 - \left(\sum V\right)^2}}$$

$$= \frac{10x141 - 2x(-2)}{\sqrt{10x140 - 2x2}\sqrt{10x176 - (-2)x(-2)}}$$

$$= \frac{1410 + 4}{\sqrt{1400 - 4}\sqrt{1760 - 4}}$$

$$= \frac{1414}{\sqrt{2451376}}$$

$$= 0.9$$

So there is a high degree of positive correlation between marks obtained in Mathematics and in Statistics.

Probable Error, denoted by *PE* (*r*) is given as

$$PE(r) = 0.6745\frac{1 - r^2}{\sqrt{N}}$$

$$PE(r) = 0.6745\frac{1-(0.9)^2}{\sqrt{10}}$$

$$PE(r) = 0.0405$$

So the value of r is highly significant.

4.3.4 SPEARMAN'S RANK CORRELATION

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as honesty, beauty, character, morality, *etc.,* which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearson's coefficient of correlation cannot be used as such. Charles Edward Spearman, a British Psychologist, developed a formula in 1904, which consists in obtaining the correlation coefficient between the ranks of N individuals in the two attributes under study.

Suppose we want to find if two characteristics A, say, intelligence and B, say, beauty are related or not. Both the characteristics are incapable of quantitative measurements but we can arrange a group of N individuals in order of merit (ranks) *w.r.t.* proficiency in the two characteristics. Let the random variables X and Y denote the ranks of the individuals in the characteristics A and B respectively. If we assume that there is no tie, *i.e.,* if no two individuals get the same rank in a characteristic then, obviously, X and Y assume numerical values ranging from *1* to N.

The Pearsonian correlation coefficient between the ranks X and Y is called the rank correlation coefficient between the characteristics A and B for the group of individuals.

Spearman's rank correlation coefficient, usually denoted by ρ(Rho) is given by the equation

$$\rho = 1 - \frac{6\sum d^2}{N(N^2-1)} \qquad \ldots\ldots\ldots\ldots(4.8)$$

Where d is the difference between the pair of ranks of the same individual in the two characteristics and N is the number of pairs.

Example 4-6

Ten entries are submitted for a competition. Three judges study each entry and list the ten in rank order. Their rankings are as follows:

Entry:	A	B	C	D	E	F	G	H	I	J
Judge J_1:	9	3	7	5	1	6	2	4	10	8
Judge J_2:	9	1	10	4	3	8	5	2	7	6
Judge J_3:	6	3	8	7	2	4	1	5	9	10

Calculate the appropriate rank correlation to help you answer the following questions:

(i) Which pair of judges agrees the most?

(ii) Which pair of judges disagrees the most?

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq.(4.8)*}

Entry	Rank by Judges			Difference in Ranks					
	J_1	J_2	J_3	$d(J_1\&J_2)$	d^2	$d(J_1\&J_3)$	d^2	$d(J_2\&J_3)$	d^2
A	9	9	6	0	0	+3	9	+3	9
B	3	1	3	+2	4	0	0	-2	4
C	7	10	8	-3	9	-1	1	+2	4
D	5	4	7	+1	1	-2	4	-3	9
E	1	3	2	-2	4	-1	1	+1	1
F	6	8	4	-2	4	+2	4	+4	16
G	2	5	1	-3	9	+1	1	+4	16
H	4	2	5	+2	4	-1	1	-3	9
I	10	7	9	+3	9	+1	1	-2	4
J	8	6	10	+2	4	-2	4	-4	16
					$\sum d^2=48$		$\sum d^2=26$		$\sum d^2=88$

$$\rho\,(J_1 \,\&\, J_2) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$=1-\frac{6\,x\,48}{10(10^2-1)}$$

$$=1-\frac{288}{990}$$

$$=1-0.29$$

$$=+0.71$$

$$\rho\,(\mathrm{J}_1\,\&\,\mathrm{J}_3) \quad =1-\frac{6\sum d^2}{N(N^2-1)}$$

$$=1-\frac{6\,x\,26}{10(10^2-1)}$$

$$=1-\frac{156}{990}$$

$$=1-0.1575$$

$$=+0.8425$$

$$\rho\,(\mathrm{J}_2\,\&\,\mathrm{J}_3) \quad =1-\frac{6\sum d^2}{N(N^2-1)}$$

$$=1-\frac{6\,x\,88}{10(10^2-1)}$$

$$=1-\frac{528}{990}$$

$$=1-0.53$$

$$=+0.47$$

So *(i)* Judges J_1 and J_3 agree the most

(ii) Judges J_2 and J_3 disagree the most

Spearman's rank correlation *Eq.(4.8)* can also be used even if we are dealing with variables, which are measured quantitatively, *i.e.* when the actual data but not the ranks relating to two variables are given. In such a case we shall have to convert the data into ranks. The highest (or the smallest) observation is given the rank 1. The next highest (or the next lowest) observation is given rank 2 and so on. It is immaterial in which way (descending or ascending) the ranks are assigned. However, the same approach should be followed for all the variables under consideration.

Example 4-7

Calculate the rank coefficient of correlation from the following data:

X:	75	88	95	70	60	80	81	50
Y:	120	134	150	115	110	140	142	100

Solution:

Calculations for Coefficient of Rank Correlation

{Using* Eq.(4.8)*}

X	Ranks R_X	Y	Ranks R_Y	$d = R_X - R_Y$	d^2
75	5	120	5	0	0
88	2	134	4	-2	4
95	1	150	1	0	0
70	6	115	6	0	0
60	7	110	7	0	0
80	4	140	3	+1	1
81	3	142	2	+1	1
50	8	100	8	0	0
					$\sum d^2 = 6$

$$\rho = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6\,x\,6}{8(8^2-1)}$$

$$= 1 - \frac{36}{504}$$

$$= 1 - 0.07$$

$$= +\,0.93$$

Hence, there is a high degree of positive correlation between *X* and *Y*

Repeated Ranks

In case of attributes if there is a tie *i.e.,* if any two or more individuals are placed together in any classification *w.r.t.* an attribute or if in case of variable data there is more than one item with the same value in either or both the series then Spearman's *Eq.(4.8)* for calculating the rank correlation coefficient breaks down, since in this case the variables *X* [the ranks of individuals in characteristic A (1st series)] and *Y* [the ranks of individuals in characteristic B (2nd series)] do not take the values from *1* to *N.*

In this case common ranks are assigned to the repeated items. These common ranks are the arithmetic mean of the ranks, which these items would have got if they were different from each other and the next item will get the rank next to the rank used in computing the common rank. For example, suppose an item is repeated at rank 4. Then the common rank to be assigned to each item is (4+5)/2, *i.e.,* 4.5 which is the average of 4 and 5, the ranks which these observations would have assumed if they were different. The next item will be assigned the rank 6. If an item is repeated thrice at rank 7, then the common rank to be assigned to each value will be (7+8+9)/3, *i.e.,* 8 which is the arithmetic mean of 7,8 and 9 *viz.,* the ranks these observations would have got if they were different from each other. The next rank to be assigned will be 10.

If only a small proportion of the ranks are tied, this technique may be applied together with *Eq.(4.8)*. If a large proportion of ranks are tied, it is advisable to apply an adjustment or a correction factor to *Eq.(4.8)*as explained below:

"In the Eq.(4.8) add the factor

$$\frac{m(m^2-1)}{12} \qquad \ldots\ldots\ldots\ldots(4.8a)$$

to $\sum d^2$*; where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the series".*

Example 4-8

For a certain joint stock company, the prices of preference shares (*X*) and debentures (*Y*) are given below:

X:	73.2	85.8	78.9	75.8	77.2	81.2	83.8
Y:	97.8	99.2	98.8	98.3	98.3	96.7	97.1

Use the method of rank correlation to determine the relationship between preference prices and debentures prices.

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq. (4.8)* and *(4.8a)*}

X	Y	Rank of X (X_R)	Rank of Y (Y_R)	$d = X_R - Y_R$	d^2
73.2	97.8	7	5	2	4
85.8	99.2	1	1	0	0
78.9	98.8	4	2	2	4
75.8	98.3	6	3.5	2.5	6.25
77.2	98.3	5	3.5	1.5	2.25
81.2	96.7	3	7	-4	16
83.8	97.1	2	6	-4	16
				$\sum d = 0$	$\sum d^2 = 48.50$

In this case, due to repeated values of *Y*, we have to apply ranking as average of 2 ranks, which could have been allotted, if they were different values. Thus ranks 3 and 4 have been allotted as 3.5 to both the values of $Y = 98.3$. Now we also have to apply correction factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, where *m* in the number of times the value is repeated, here $m = 2$.

$$\rho \quad = \quad \frac{6\left[\sum d^2 + \frac{m\left(m^2-1\right)}{2}\right]}{N(N^2-1)}$$

$$= \quad \frac{6\left[48.5 + \frac{2(4-1)}{12}\right]}{7(7^2-1)}$$

$$= \quad 1- \frac{6\,x\,49}{7\,x\,48}$$

$$= \quad 0.125$$

Hence, there is a very low degree of positive correlation, probably no correlation, between preference share prices and debenture prices.

Remarks on Spearman's Rank Correlation Coefficient

1. We always have $\sum d = 0$, which provides a check for numerical calculations.

2. Since Spearman's rank correlation coefficient, ρ, is nothing but Karl Pearson's correlation coefficient, *r*, between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.

3. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure, which is distribution free (or non-parametric). Spearman's ρ is such a distribution free measure, since no strict assumption are made about the from of the population from which sample observations are drawn.

4. Spearman's formula is easy to understand and apply as compared to Karl Pearson's formula. The values obtained by the two formulae, *viz* Pearsonian *r* and Spearman's ρ are generally different. The difference arises due to the fact that when ranking is used instead of full set of observations, there is always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics, which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.

6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution. For N >30, this formula should not be used unless the ranks are given.

4.3.5 CONCURRENT DEVIATION METHOD

This is a casual method of determining the correlation between two series when we are not very serious about its precision. This is based on the signs of the deviations (*i.e.* the direction of the change) of the values of the variable from its preceding value and does not take into account the exact magnitude of the values of the variables. Thus we put a plus (+) sign, minus (-) sign or equality (=) sign for the deviation if the value of the variable is greater than, less than or equal to the preceding value respectively. The deviations in the values of two variables are said to be concurrent if they have the same sign (either both deviations are positive or both are negative or both are equal). The formula used for computing correlation coefficient r_c by this method is given by

$$r_c = \pm\sqrt{\pm\left(\frac{2c-N}{N}\right)} \qquad \ldots\ldots\ldots\ldots(4.9)$$

Where c is the number of pairs of concurrent deviations and N is the number of pairs of deviations. If ($2c$-N) is positive, we take positive sign in and outside the square root in *Eq. (4.9)* and if ($2c$-N) is negative, we take negative sign in and outside the square root in *Eq. (4.9).*

Remarks: *(i)* It should be clearly noted that here N is not the number of pairs of observations but it is the number of pairs of deviations and as such it is one less than the number of pairs of observations.

(ii) Coefficient of concurrent deviations is primarily based on the following principle:

"If the short time fluctuations of the time series are positively correlated or in other words, if their deviations are concurrent, their curves would move in the same direction and would indicate positive correlation between them"

Example 4-9

Calculate coefficient of correlation by the concurrent deviation method

Supply:	112	125	126	118	118	121	125	125	131	135
Price:	106	102	102	104	98	96	97	97	95	90

Solution:

Calculations for Coefficient of Concurrent Deviations

{Using *Eq. (4.9)*}

Supply (X)	Sign of deviation from preceding value (X)	Price (Y)	Sign of deviation preceding value (Y)	Concurrent deviations
112		106		
125	+	102	-	
126	+	102	=	
118	-	104	+	
118	=	98	-	
121	+	96	-	
125	+	97	+	+(c)
125	=	97	=	= (c)
131	+	95	-	
135	+	90	-	

We have

Number of pairs of deviations, N =10 – 1 = 9

c = Number of concurrent deviations

= Number of deviations having like signs

= 2

Coefficient of correlation by the method of concurrent deviations is given by:

$$r_c = \underline{+}\sqrt{\underline{+}\left(\frac{2c-N}{N}\right)}$$

$$r_c = \underline{+}\sqrt{\underline{+}\left(\frac{2x2-9}{9}\right)}$$

$$r_c = \underline{+}\sqrt{\underline{+}(-0.5556)}$$

Since $2c - N$ = -5 (negative), we take negative sign inside and outside the square root

$$r_c = -\sqrt{-(-0.5556)}$$

$$r_c = -\sqrt{0.5556}$$

$$r_c = -0.7$$

Hence there is a fairly good degree of negative correlation between supply and price.

4.4 LIMITATIONS OF CORRELATION ANALYSIS

As mentioned earlier, correlation analysis is a statistical tool, which should be properly used so that correct results can be obtained. Sometimes, it is indiscriminately used by management, resulting in misleading conclusions. We give below some *errors* frequently made in the use of correlation analysis:

1. Correlation analysis cannot determine cause-and-effect relationship. One should not assume that a change in *Y* variable is caused by a change in *X* variable unless one is reasonably sure that one variable is the cause while the other is the effect. Let us take an example. .

   Suppose that we study the performance of students in their graduate examination and their earnings after, say, three years of their graduation. We may find that these two variables are highly and positively related. At the same time, we must not forget that both the variables might have been influenced by some other factors such as quality of teachers, economic and social status of parents, effectiveness of the interviewing process and so forth. If the data on these factors are available, then it is worthwhile to use multiple correlation analysis instead of bivariate one.

2. Another mistake that occurs frequently is on account of misinterpretation of the coefficient of correlation. Suppose in one case $r = 0.7$, it will be wrong to interpret that correlation explains 70 percent of the total variation in *Y.* The error can be seen easily when we calculate the coefficient of determination. Here, the coefficient of

determination r^2 will be 0.49. This means that only 49 percent of the total variation in Y is explained.

Similarly, the coefficient of determination is misinterpreted if it is also used to indicate causal relationship, that is, the percentage of the change in one variable is due to the change in another variable.

3. Another mistake in the interpretation of the coefficient of correlation occurs when one concludes a positive or negative relationship even though the two variables are actually unrelated. For example, the age of students and their score in the examination have no relation with each other. The two variables may show similar movements but there does not seem to be a common link between them.

To sum up, one has to be extremely careful while interpreting coefficient of correlation. Before one concludes a causal relationship, one has to consider other relevant factors that might have any influence on the dependent variable or on both the variables. Such an approach will avoid many of the pitfalls in the interpretation of the coefficient of correlation. It has been rightly said that the ***coefficient of correlation is not only one of the most widely used, but also one of the widely abused statistical measures.***

Objectives: **The overall objective of this lesson is to give you an understanding of linear regression, there by enabling you to understand the importance and also the limitations of regression analysis.**

Structure

...if we find any association between two or more variables, we might be interested in estimating the value of one variable for known value(s) of another variable(s)

5.1 INTRODUCTION

In business, several times it becomes necessary to have some forecast so that the management can take a decision regarding a product or a particular course of action. In order to make a forecast, one has to ascertain some relationship between two or more variables relevant to a particular situation. For example, a company is interested to know how far the demand for television sets will increase in the next five years, keeping in mind the growth of population in a certain town. Here, it clearly assumes that the increase in population will lead to an increased demand for television sets. Thus, to determine the nature and extent of relationship between these two variables becomes important for the company.

In the preceding lesson, we studied in some depth linear correlation between two variables. Here we have a similar concern, the association between variables, except that we develop it further in two respects. *First,* we learn how to build statistical models of relationships between the variables to have a better understanding of their features. *Second,* we extend the models to consider their use in forecasting.

For this purpose, we have to use the technique - ***regression analysis*** **-** which forms the subject-matter of this lesson.

5.2 WHAT IS REGRESSION?

In 1889, Sir Francis Galton, a cousin of Charles Darwin published a paper on heredity, *"Natural Inheritance".* He reported his discovery that sizes of seeds of sweet pea plants appeared to "revert" or "regress", to the mean size in successive generations. He also reported results of a study of the relationship between heights of fathers and heights of their sons. A straight line was fit to the data pairs: *height of father versus height of son.* Here, too, he found a "regression to mediocrity" The heights of the sons represented a movement away from their fathers, towards the average height. We credit Sir Galton with the idea of statistical regression.

While most applications of regression analysis may have little to do with the "regression to the mean" discovered by Galton, the term **"regression"** remains. It now refers to ***the statistical technique of modeling the relationship between two or more variables.*** In general sense, regression analysis means the estimation or prediction of the unknown value of one variable from the known value(s) of the other variable(s). It is one of the most important and widely used statistical techniques in almost all sciences - natural, social or physical.

In this lesson we will focus only on **simple regression** –linear regression involving only two variables: a dependent variable and an independent variable. Regression analysis for studying more than two variables at a time is known as **multiple regressions.**

5.2.1 INDEPENDENT AND DEPENDENT VARIABLES

Simple regression involves only two variables; one variable is predicted by another variable. *The variable to be predicted* is called the **dependent variable**. *The predictor* is called the **independent variable,** or *explanatory variable.* For example, when we are trying to predict the demand for television sets on the basis of population growth, we are using the demand for television sets as the dependent variable and the population growth as the independent or predictor variable.

The decision, as to which variable is which sometimes, causes problems. Often the choice is obvious, as in case of demand for television sets and population growth because it would make no sense to suggest that population growth could be dependent on TV demand! The population growth has to be the independent variable and the TV demand the dependent variable.

If we are unsure, here are some points that might be of use:

- if we have control over one of the variables then that is the independent. For example, a manufacturer can decide how much to spend on advertising and expect his sales to be dependent upon how much he spends
- it there is any lapse of time between the two variables being measured, then the latter must depend upon the former, it cannot be the other way round
- if we want to predict the values of one variable from your knowledge of the other variable, the variable to be predicted must be dependent on the known one

5.3 LINEAR REGRESSION

The task of bringing out linear relationship consists of developing methods of fitting a straight line, or a regression line as is often called, to the data on two variables.

The line of Regression is the graphical or relationship representation of the best estimate of one variable for any given value of the other variable. The nomenclature of the line depends on the independent and dependent variables. If *X* and *Y* are two variables of which relationship is to be indicated, a line that gives best estimate of *Y* for any value of *X,* it is called ***Regression line of Y on X.*** If the dependent variable changes to *X,* then best estimate of *X* by any value of *Y* is called ***Regression line of X on Y.***

5.3.1 REGRESSION LINE OF *Y* ON *X*

For purposes of illustration as to how a straight line relationship is obtained, consider the sample paired data on sales of each of the *N* = 5 months of a year and the marketing expenditure incurred in each month, as shown in Table 5-1

Table 5-1

Month	Sales (Rs lac)	Marketing Expenditure (Rs thousands)

	Y	***X***
April	14	10
May	17	12
June	23	15
July	21	20
August	25	23

Let *Y*, the sales, be the dependent variable and *X*, the marketing expenditure, the independent variable. We note that for each value of independent variable *X*, there is a specific value of the dependent variable *Y*, so that each value of *X* and *Y* can be seen as paired observations.

5.3.1.1 Scatter Diagram

Before obtaining a straight-line relationship, it is necessary to discover whether the relationship between the two variables is linear, that is, the one which is best explained by a straight line. A good way of doing this is to plot the data on *X* and *Y* on a graph so as to yield a scatter diagram, as may be seen in Figure 5-1. A careful reading of the scatter diagram reveals that:

- the overall tendency of the points is to move upward, so the relationship is positive
- the general course of movement of the various points on the diagram can be best explained by a straight line
- there is a high degree of correlation between the variables, as the points are very close to each other

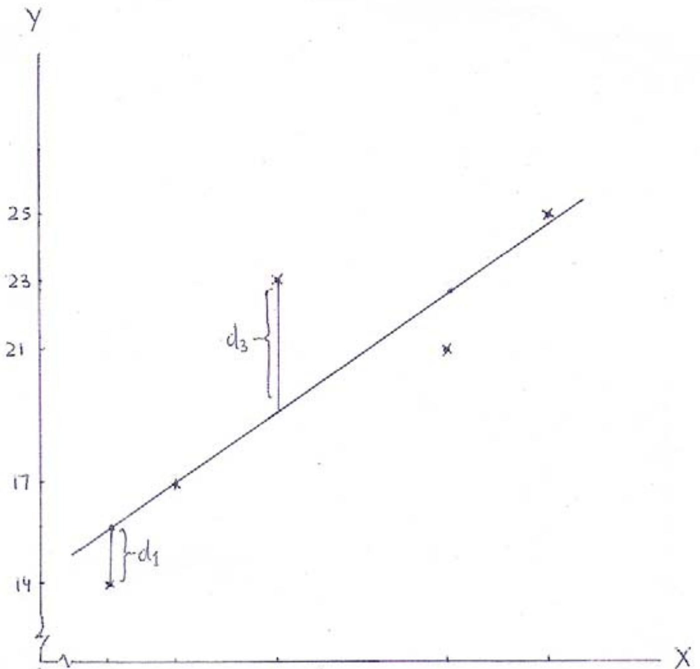


Figure 5-1 Scatter Diagram with Line of Best Fit

5.3.1.2 Fitting a Straight Line on the Scatter Diagram

If the movement of various points on the scatter diagram is best described by a straight line, the next step is to fit a straight line on the scatter diagram. It has to be so fitted that on the whole it lies as close as possible to every point on the scatter diagram. The necessary requirement for meeting this condition being that ***the sum of the squares of the vertical deviations of the observed Y values from the straight line is minimum.***

As shown in Figure 5-1, if d_1, d_2,..., d_N are the vertical deviations' of observed Y values from the straight line, fitting a straight line requires that

$$d_1^2 + d_2^2 + ... + d_N^2 = \sum_{j=1}^{N} d_j^2$$

is the minimum. The deviations d_j have to be squared to avoid negative deviations canceling out the positive deviations. Since a straight line so fitted best approximates all the points on the scatter diagram, it is better known as the best approximating line or the ***line of best fit***. A line of best fit can be fitted by means of:

1. Free hand drawing method, and
2. Least square method

<u>Free Hand Drawing:</u>

Free hand drawing is the simplest method of fitting a straight line. After a careful inspection of the movement and spread of various points on the scatter diagram, a straight line is drawn through these points by using a transparent ruler such that on the whole it is closest to every point. A straight line so drawn is particularly useful when future approximations of the dependent variable are promptly required.

Whereas the use of free hand drawing may yield a line nearest to the line of best fit, the major drawback is that the slope of the line so drawn varies from person to person because of the influence of subjectivity. Consequently, the values of the dependent variable estimated on the basis of such a line may not be as accurate and precise as those based on the line of best fit.

<u>Least Square Method:</u>

The least square method of fitting a line of best fit requires minimizing the sum of the squares of vertical deviations of each observed *Y* value from the fitted line. These deviations, such as d_1 and d_3, are shown in Figure 5-1 and are given by Y - Y_c, where *Y* is the observed value and Y_c the corresponding computed value given by the fitted line

$$Y_c = a + bX_i \qquad \qquad \text{…………}(5.1)$$

for the i^{th} value of *X*.

The straight line relationship in *Eq.(5.1),* is stated in terms of two constants *a* and *b*

- The constant *a* is the *Y*-intercept; it indicates the height on the vertical axis from where the straight line originates, representing the value of *Y* when *X* is zero.
- Constant *b* is a measure of the slope of the straight line; it shows the absolute change in *Y* for a unit change in *X*. As the slope may be positive or negative, it indicates the nature of relationship between *Y* and *X*. Accordingly, *b* is also known as the regression coefficient of *Y* on *X*.

Since a straight line is completely defined by its intercept *a* and slope *b,* the task of fitting the same reduces only to the computation of the values of these two constants. Once these two values are known, the computed Y_c values against each value of *X* can be easily obtained by substituting *X* values in the linear equation.

In the method of least squares the values of *a* and *b* are obtained by solving simultaneously the following pair of normal equations

$$\sum Y = aN + b\sum X \qquad \ldots\ldots\ldots\ldots(5.2)$$

$$\sum XY = a\sum X + b\sum X^2 \qquad \ldots\ldots\ldots\ldots(5.2)$$

The value of the expressions - $\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$ can be obtained from the given observations and then can be substituted in the above equations to obtain the value of *a* and *b*. Since simultaneous solving the two normal equations for *a* and *b* may quite often be cumbersome and time consuming, the two values can be directly obtained as

$$a = \overline{Y} - b\overline{X} \qquad \ldots\ldots\ldots\ldots(5.3)$$

and

$$b = \frac{N\sum XY - \sum X\sum Y}{N\sum X^2 - \left(\sum X\right)^2} \qquad \ldots\ldots\ldots\ldots(5.4)$$

Note: Eq. (5.3) is obtained simply by dividing both sides of the first of Eqs. (5.2) by N and Eq.(5.4) is obtained by substituting ($\overline{Y} - b\overline{X}$) in place of a in the second of Eqs. (5.2)

Instead of directly computing *b*, we may first compute value of *a* as

$$a = \frac{\sum Y\sum X^2 - \sum X\sum XY}{N\sum X^2 - \left(\sum X\right)^2} \qquad \ldots\ldots\ldots\ldots(5.5)$$

and

$$b = \frac{\overline{Y} - a}{\overline{X}} \qquad \ldots\ldots\ldots\ldots(5.6)$$

Note: Eq. (5.5) is obtained by substituting $\frac{N\sum XY - \sum X\sum Y}{N\sum X^2 - \left(\sum X\right)^2}$ *for b in Eq. (5.3) and Eq. (5.6) is obtained simply by rearranging Eq. (5.3)*

Table 5-2
***Computation of* a *and* b**

Y	*X*	*XY*	X^2	Y^2

14	10	140	100	196
17	12	204	144	289
23	15	345	225	529
21	20	420	400	441
25	23	575	529	625
$\sum Y = 100$	$\sum X = 80$	$\sum XY = 1684$	$\sum X^2 = 1398$	$\sum Y^2 = 2080$

So using *Eqs. (5.5)* and *(5.4)*

$$a = \frac{100x1398 - 80x1684}{5x1398 - (80)^2}$$

$$= \frac{139800 - 134720}{6990 - 6400}$$

$$= \frac{5080}{590}$$

$$= 8.6101695$$

and

$$b = \frac{5x1684 - 80x100}{5x1398 - (80)^2}$$

$$= \frac{8420 - 8000}{6990 - 6400}$$

$$= \frac{420}{590}$$

$$= 0.7118644$$

Now given $a = 8.61$ and $b = 0.71$

The regression *Eq.(5.1)* takes the form

$$Y_c = 8.61 + 0.71X \qquad \ldots\ldots\ldots\ldots(5.1a)$$

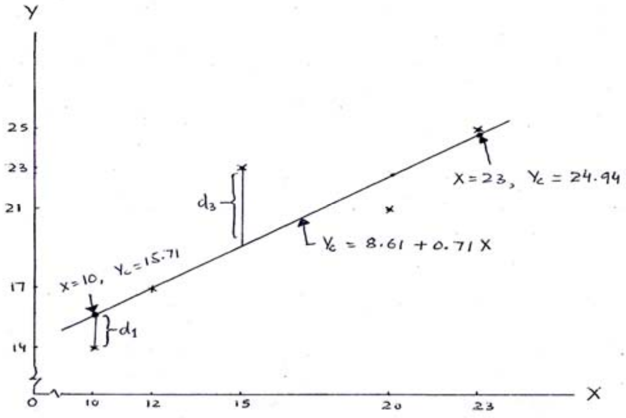


Figure 5-2 Regression Line of *Y* on *X*

Then, to fit the line of best fit on the scatter diagram, only two computed Y_c values are needed. These can be easily obtained by substituting any two values of *X* in *Eq. (5.1a).* When these are plotted on the diagram against their corresponding values of *X*, we get two points, by joining which (by means of a straight line) gives us the required line of best fit, as shown in Figure 5-2

Some Important Relationships

We can have some important relationships for data analysis, involving other measures such as $\overline{X}$, $\overline{Y}$, S_x, S_y and the correlation coefficient r_{xy}.

Substituting $\overline{Y} - b\overline{X}$ [from *Eq.(5.3)*] for *a* in *Eq.(5.1)*

$$Y_c = (\overline{Y} - b\overline{X}) + bX$$

$$\text{or} \qquad Y_c - \overline{Y} = b(X - \overline{X}) \qquad \ldots\ldots\ldots\ldots(5.7)$$

Dividing the numerator and denominator of *Eq.(5.4)* by N^2, we get

$$b = \frac{\frac{\sum XY}{N} - \left(\frac{\sum X}{N}\right)\left(\frac{\sum Y}{N}\right)}{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2}$$

$$\text{or} \qquad b = \frac{\frac{\sum XY}{N} - \overline{X}\,\overline{Y}}{S_x^2}$$

$$\text{or} \qquad b = \frac{Cov(X,Y)}{S_x^2} \qquad \ldots\ldots\ldots\ldots(5.8)$$

We know, coefficient of correlation, r_{xy} is given by

$$r_{xy} = \frac{Cov(X,Y)}{S_x\,S_y}$$

or $Cov(X,Y) = r_{xy} S_x \, S_y$

So *Eq. (5.8)* becomes

$$b = r_{xy} \frac{S_x S_y}{S_x^2}$$

$$b = r_{xy} \frac{S_y}{S_x} \qquad \text{…………}(5.9)$$

Substituting $r_{xy} \frac{S_y}{S_x}$ for *b* in *Eq.(5.7),* we get

$$Y_c - \overline{Y} = r_{xy} \frac{S_y}{S_x} \; (X - \overline{X}) \qquad \text{…………}(5.10)$$

These are important relationships for data analysis.

5.3.1.3 Predicting an Estimate and its Preciseness

The main objective of regression analysis is to know the nature of relationship between two variables and to use it for predicting the most likely value of the dependent variable corresponding to a given, known value of the independent variable. This can be done by substituting in *Eq.(5.1a)* any known value of *X* corresponding to which the most likely estimate of *Y* is to be found.

For example, the estimate of *Y* (*i.e.* Y_c), corresponding to $X = 15$ is

$$\begin{aligned} Y_c &= 8.61 + 0.71(15) \\ &= 8.61 + 10.65 \\ &= 19.26 \end{aligned}$$

It may be appreciated that an estimate of *Y* derived from a regression equation will not be exactly the same as the *Y* value which may actually be observed. The difference between estimated Y_c values and the corresponding observed *Y* values will depend on the extent of scatter of various points around the line of best fit.

The closer the various paired sample points *(Y, X)* clustered around the line of best fit, the smaller the difference between the estimated Y_c and observed Y values, and vice-versa. On the whole, the lesser the scatter of the various points around, and the lesser the vertical distance by which these deviate from the line of best fit, the more likely it is that an estimated Y_c value is close to the corresponding observed Y value.

The estimated Y_c values will coincide the observed Y values only when all the points on the scatter diagram fall in a straight line. If this were to be so, the sales for a given marketing expenditure could have been estimated with l00 percent accuracy. But such a situation is too rare to obtain. Since some of the points must lie above and some below the straight line, perfect prediction is practically non-existent in the case of most business and economic situations.

This means that the estimated values of one variable based on the known values of the other variable are always bound to differ. The smaller the difference, the greater the precision of the estimate, and vice-versa. Accordingly, the preciseness of an estimate can be obtained only through a measure of the magnitude of error in the estimates, called the ***error of estimate***.

5.3.1.4 Error of Estimate

A measure of the error of estimate is given by the standard error of estimate of *Y on X*, denoted as S_{yx} and defined as

$$S_{yx} = \sqrt{\frac{\sum (Y - Y_c)^2}{N}} \qquad \text{…………}(5.11)$$

S_{yx} measures the average absolute amount by which observed Y values depart from the corresponding computed Y_c values.

Computation of S_{yx} becomes little cumbersome where the number of observations N is large. In such cases S_{yx} may be computed directly by using the equation:

$$S_{yx} = \sqrt{\frac{\sum Y^2 - a\left(\sum Y\right) - b\sum XY}{N}} \qquad \ldots\ldots\ldots\ldots(5.12)$$

By substituting the values of $\sum Y^2$, $\sum Y$, and $\sum XY$ from the Table 5-2, and the calculated values of *a* and *b*

We have

$$\begin{aligned} S_{yx} &= \sqrt{\frac{2080 - 8.61x100 - 0.71x1684}{5}} \\ &= \sqrt{\frac{2080 - 861 - 1195.64}{5}} \\ &= \sqrt{\frac{23.36}{5}} \\ &= \sqrt{4.67} \\ &= 2.16 \end{aligned}$$

<u>Interpretations of S_{yx}</u>

A careful observation of how the standard error of estimate is computed reveals the following:

1. S_{yx} is a concept statistically parallel to the standard deviation S_y . The only difference between the two being that the standard deviation measures the dispersion around the mean; the standard error of estimate measures the dispersion around the regression line. Similar to the property of arithmetic mean, the sum of the deviations of different *Y* values from their corresponding estimated Y_c values is equal to zero. That is

   $\sum(Y_i - \overline{Y}) = \sum (Y_i - Y_c) = 0$ where $i = 1, 2, ..., N$.

2. S_{yx} tells us the amount by which the estimated Y_c values will, on an average, deviate from the observed *Y* values. Hence it is an estimate of the average amount of error in the estimated Y_c values. The actual error (the residual of *Y* and Y_c) may, however, be smaller or larger than the average error. Theoretically, these errors follow a normal distribution. Thus, assuming that $n \geq 30$, $Y_c \pm 1.S_{yx}$ means that 68.27% of the estimates

based on the regression equation will be within 1.S_{yx} Similarly, Y_c ± 2.S_{yx} means that 95.45% of the estimates will fall within 2.S_{yx}

Further, for the estimated value of sales against marketing expenditure of Rs 15 thousand being Rs 19.26 lac, one may like to know how good this estimate is. Since S_{yx} is estimated to be Rs 2.16 lac, it means there are about 68 chances (68.27) out of 100 that this estimate is in error by not more than Rs 2.16 lac above or below Rs 19.26 lac. That is, there are 68% chances that actual sales would fall between (19.26 - 2.16) = Rs 17.10 lac and (19.26 + 2.16) = Rs 21.42 lac.

3. Since S_{yx} measures the closeness of the observed *Y* values and the estimated Y_c values, it also serves as a measure of the reliability of the estimate. Greater the closeness between the observed and estimated values of *Y,* the lesser the error and, consequently, the more reliable the estimate. And vice-versa.
4. Standard error of estimate S_{yx} can also be seen as a measure of correlation insofar as it expresses the degree of closeness of scatter of observed *Y* values about the regression line. The closer the observed *Y* values scattered around the regression line, the higher the correlation between the two variables.

   A major difficulty in using S_{yx} as a measure of correlation is that it is expressed in the same units of measurement as the data on the dependent variable. This creates problems in situations requiring comparison of two or more sets of data in terms of correlation. It is mainly due to this limitation that the standard error of estimate is not generally used as a measure of correlation. However, it does serve as the basis of evolving the coefficient of determination, denoted as r^2, which provides an alternate method of obtaining a measure of correlation.

5.3.2 REGRESSION LINE OF *X* ON *Y*

So far we have considered the regression of *Y* on *X*, in the sense that *Y* was in the role of dependent and *X* in the role of an independent variable. In their reverse position, such that *X* is now the dependent and *Y* the independent variable, we fit a line of regression of *X* on *Y*. The regression equation in this case will be

$$X_c = a' + b'Y \qquad \text{…………(5.13)}$$

Where X_c denotes the computed values of *X* against the corresponding values of *Y*. *a*' is the *X*-intercept and *b*' is the slope of the straight line.

Two normal equations to solve *a*'and *b*' are

$$\sum X = a'N + b'\sum Y \qquad \text{…………(5.14)}$$

$$\sum XY = a'\sum Y + b'\sum Y^2 \qquad \text{…………(5.14)}$$

The value of *a*' and *b*' can also be obtained directly

$$a' = \overline{X} - b'\overline{Y} \qquad \text{…………(5.15)}$$

and

$$b' = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - \left(\sum Y\right)^2} \qquad \text{…………(5.16)}$$

or

$$a' = \frac{\sum X \sum Y^2 - \sum Y \sum XY}{N\sum Y^2 - \left(\sum Y\right)^2} \qquad \text{…………(5.17)}$$

and

$$b' = \frac{\overline{X} - a'}{\overline{Y}} \qquad \text{…………(5.18)}$$

$$b' = \frac{Cov(Y, X)}{S_y^2} \qquad \text{…………(5.19)}$$

$$b' = r_{yx}\frac{S_x}{S_y} \qquad \text{…………(5.20)}$$

So, Regression equation of *X* on *Y* may also be written as

$$X_c - \overline{X} = b' (Y - \overline{Y}) \qquad \text{…………(5.21)}$$

$$X_c - \overline{X} = r_{yx} \frac{S_x}{S_y} (Y - \overline{Y}) \qquad \text{…………(5.22)}$$

As before, once the values of *a*' and *b*' have been found, their substitution in *Eq.(5.13)* will enable us to get an estimate of *X* corresponding to a known value of *Y*

Standard Error of estimate of *X* on *Y i.e.* S_{xy} will be

$$S_{xy} = \sqrt{\frac{(X - X_c)^2}{N}} \qquad \text{…………(5.23)}$$

or

$$S_{xy} = \sqrt{\frac{\sum X^2 - a'\sum X - b'\sum XY}{N}} \qquad \text{…………(5.24)}$$

For example, if we want to estimate the marketing expenditure to achieve a sale target of Rs 40 lac, we have to obtain regression line of *X* on *Y i. e.*

$$X_c = a' + b'Y$$

So using *Eqs. (5.17)* and *(5.16),* and substituting the values of $\sum X, \sum Y^2, \sum Y$ and $\sum XY$ from Table 5-2, we have

$$a' = \frac{80x2080 - 100x1684}{5x2080 - (100)^2}$$

$$= \frac{166400 - 168400}{10400 - 10000}$$

$$= \frac{-2000}{400}$$

$$= -5.00$$

and

$$b' = \frac{5x1684 - 80x100}{5x2080 - (100)^2}$$

$$= \frac{8420 - 8000}{10400 - 10000}$$

$= \frac{420}{400}$

$= 1.05$

Now given that *a*'= -5.00 and *b*'=1.05, Regression equation *(5.13)* takes the form

$X_c = -5.00 + 1.05Y$

So when *Y* = 40(Rs lac), the corresponding *X* value is

$X_c = -5.00 + 1.05x40$

$= -5 + 42$

$= 37$

That is to achieve a sale target of Rs 40 lac, there is a need to spend Rs 37 thousand on marketing.

5.4 PROPERTIES OF REGRESSION COEFFICIENTS

As explained earlier, the slope of regression line is called the regression coefficient. It tells the effect on dependent variable if there is a unit change in the independent variable. Since for a paired data on *X* and *Y* variables, there are two regression lines: regression line of *Y* on *X* and regression line of *X* on *Y*, so we have two regression coefficients:

a. Regression coefficient of *Y* on *X*, denoted by b_{yx} [*b* in *Eq.(5.1)*]

b. Regression coefficient of *X* on *Y*, denoted by b_{xy} [*b*' in *Eq.(5.13)*]

The following are the important properties of regression coefficients that are helpful in data analysis

1. The value of both the regression coefficients cannot be greater than 1. However, value of both the coefficients can be below 1 or at least one of them must be below 1, so that the square root of the product of two regression coefficients must lie in the limit ±1.
2. Coefficient of correlation is the geometric mean of the regression coefficients, *i.e.*

$$r = \pm\sqrt{b.b'} \qquad \ldots\ldots\ldots\ldots(5.25)$$

The signs of both the regression coefficients are the same, and so the value of *r* will also have the same sign.

3. The mean of both the regression coefficients is either equal to or greater than the coefficient of correlation, *i.e.*

$$\frac{b+b'}{2} \geq r$$

3. Regression coefficients are independent of change of origin but not of change of scale. Mathematically, if given variables *X* and *Y* are transformed to new variables *U* and *V* by change of origin and scale, *i. e.*

$$U = \frac{X-A}{h} \quad \text{and} \quad V = \frac{Y-B}{k}$$

Where *A, B, h* and *k* are constants, *h > 0, k > 0* then

Regression coefficient of *Y* on *X* = *k/h* (Regression coefficient of *V* on *U*)

$$b_{yx} = \frac{k}{h} b_{vu}$$

and

Regression coefficient of *X* on *Y* = *h/k* (Regression coefficient of *U* on *V*)

$$b_{xy} = \frac{h}{k} b_{uv}$$

5. Coefficient of determination is the product of both the regression coefficients *i.e.*

$$r^2 = b.b'$$

5.5 REGRESSION LINES AND COEFFICIENT OF CORRELATION

The two regression lines indicate the nature and extent of correlation between the variables.

The two regression lines can be represented as

$$Y - \overline{Y} = r\frac{S_y}{S_x}(X - \overline{X}) \qquad \text{and} \qquad X - \overline{X} = r\frac{S_x}{S_y}(Y - \overline{Y})$$

We can write the slope of these lines, as

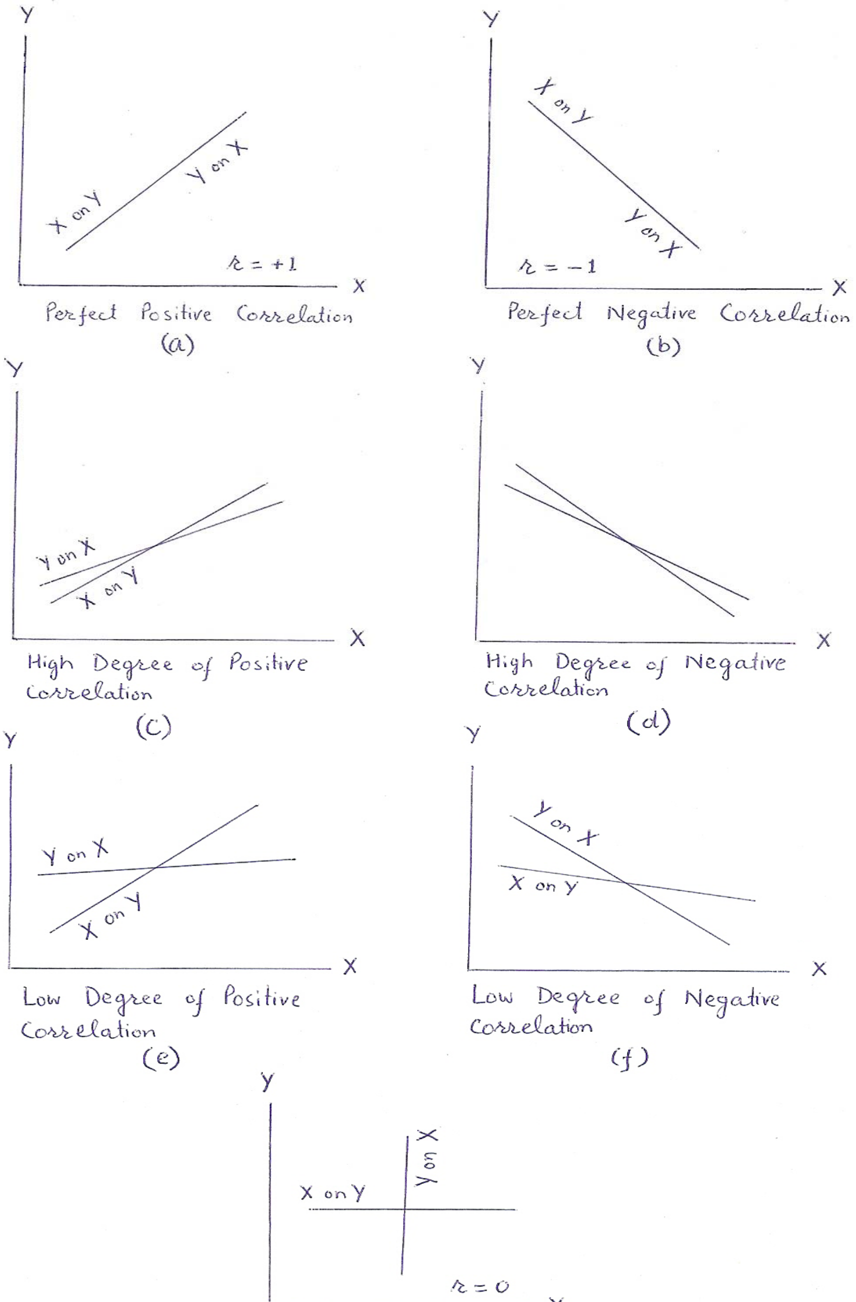
$$b = r\frac{S_y}{S_x} \qquad \text{and} \qquad b' = r\frac{S_x}{S_y}$$

If θ is the angle between these lines, then

$$\tan\theta = \frac{b-b'}{1+bb'}$$

$$= \frac{S_x S_y}{S_x^2 + S_y^2}\left(\frac{r^2-1}{r}\right)$$

$$\text{or } \theta = \tan^{-1}\left[\frac{S_x S_y}{S_x^2 + S_y^2}\left(\frac{r^2-1}{r}\right)\right] \qquad \text{…………(5.26)}$$



Perfect Positive Correlation
(a)

Perfect Negative Correlation
(b)

High Degree of Positive Correlation
(c)

High Degree of Negative Correlation
(d)

Low Degree of Positive Correlation
(e)

Low Degree of Negative Correlation
(f)

Figure 5-3 Regression Lines and Coefficient of Correlation

Eq. (5.26) reveals the following:

- In case of perfect positive correlation (r = +1) and in case of perfect negative correlation (r = -1), $\theta = 0$, so the two regression lines will coincide, *i.e.* we have only one line, see (a) and (b) in Figure 5-3.

  The farther the two regression lines from each other, lesser will be the degree of correlation and nearer the two regression lines, more will be the degree of correlation, see (c) and (d) in Figure 5-3.

- If the variables are independent *i.e.* $r = 0$, the lines of regression will cut each other at right angle. See (g) in Figure 5-3.

Note : Both the regression lines cut each other at mean value of X and mean value of Y i.e. at $\overline{X}$ *and* $\overline{Y}$.

5.6 COEFFICIENT OF DETERMINATION

Coefficient of determination gives the percentage variation in the dependent variable that is accounted for by the independent variable. In other words, the coefficient of determination gives the ratio of the explained variance to the total variance. The coefficient of determination is given by the square of the correlation coefficient, *i.e.* r^2. Thus,

Coefficient of determination

$$r^2 = \frac{Explained\ Variance}{Total\ Variance}$$

$$r^2 = \frac{\sum\left(Y_c - \overline{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2} \qquad \ldots\ldots\ldots\ldots(5.27)$$

We can calculate another coefficient K^2, known as coefficient of Non-Determination, which is the ratio of unexplained variance to the total variance.

$$K^2 = \frac{Unexplained\ Variance}{Total\ Variance}$$

$$K^2 = \frac{\sum (Y - Y_c)^2}{\sum (Y - \overline{Y})^2} \qquad \ldots\ldots\ldots\ldots(5.28)$$

$$K^2 = 1 - \frac{Explained\ Variance}{Total\ Variance}$$

$$= 1 - r^2 \qquad \ldots\ldots\ldots\ldots(5.29)$$

The square root of the coefficient of non-determination, *i.e.* K gives the coefficient of alienation

$$K = \pm\sqrt{1 - r^2} \qquad \ldots\ldots\ldots\ldots(5.30)$$

Relation Between S_{yx} and r:

A simple algebraic operation on *Eq. (5.30)* brings out some interesting points about the relation between S_{yx} and r. Thus, since

$$\sum (Y - Y_c)^2 = N\,S_{yx}^2 \qquad \text{and} \qquad \sum (Y - \overline{Y})^2 = N\,S_y^2$$

So we have coefficient of Non-determination

$$K^2 = \frac{\sum (Y - Y_c)^2}{\sum (Y - \overline{Y})^2}$$

$$K^2 = \frac{N\,S_{yx}^2}{N\,S_y^2}$$

$$= \frac{S_{yx}^2}{S_y^2}$$

So $$1 - r^2 = \frac{S_{yx}^2}{S_y^2}$$

or $$\frac{S_{yx}}{S_y} = \sqrt{1 - r^2} \qquad \ldots\ldots\ldots\ldots(5.31)$$

If coefficient of correlation, *r*, is defined as the under root of the coefficient of determination

$$r = \sqrt{r^2}$$

$$r^2 = \quad 1 - \frac{S_{yx}^2}{S_y^2}$$

$$r = \sqrt{1 - \frac{S_{yx}}{S_y^2}} \qquad \ldots\ldots\ldots\ldots(5.32)$$

On carefully observing *Eq. (5.32),* it will be noticed that the ratio S_{yx}/S_y will be large if the coefficient of determination is small, and it will be small when the coefficient of determination is large. Thus

✓ if $r^2 = r = 0$, $S_{yx}/S_y = 1$, which means that $S_{yx} = S_y$.

✓ if $r^2 = r = 1$, $S_{yx}/S_y = 0$, which means that $S_{yx} = 0$.

✓ when $r = 0.865$, $S_{yx} = 0.427\ S_y$ means that S_{yx} is 42.7% of S_y.

Eq. (5.32) also implies that S_{yx} is generally less than S_y. The two can at the most be equal, but only in the extreme situation when $r = 0$.

<u>Interpretations of r^2:</u>

1. Even though the coefficient of determination, whose under root measures the degree of correlation, is based on $S_{yx},$*;* it is expressed as 1 - (S_{yx}/S_y). As it is a dimensionless pure number, the unit in which S_{yx} is measured becomes irrelevant. This facilitates comparison between the two sets of data in terms of their coefficient of determination r^2 (or the coefficient of correlation *r).* This was not possible in terms of $S_{y\ x}$ as the units of measurement could be different.
2. The value of r^2 can range between 0 and 1. When $r^2 = 1$, all the points on the scatter diagram fall on the regression line and the entire variations are explained by the straight line. On the other hand, when $r^2 = 0$, none of the points on the scatter diagram falls on the regression line, meaning thereby that there is no relationship between the two variables. However, being always non-negative coefficient of determination does

not tell us about the direction of the relationship (whether it is positive or negative) between the two variables.

3. When r^2 = 0.7455 (or any other value), 74.55% of the total variations in sales are explained by the marketing expenditure used. What remains is the coefficient of non-determination K^2 (= 1 - r^2) = 0.2545. It means 25.45% of the total variations remain unexplained, which are due to factors other than the changes in the marketing expenditure.
4. r^2 provides the necessary link between regression and correlation which are the two related aspects of a single problem of the analysis of relationship between two variables. Unlike regression, correlation quantifies the degrees of relationship between the variables under study, without making a distinction between the dependent and independent ones. Nor does it, therefore, help in predicting the value of one variable for a given value of the other.
5. The coefficient of correlation overstates the degree of relationship and it's meaning is not as explicit as that of the coefficient of determination. The coefficient of correlation r = 0.865, as compared to r^2 = 0.7455, indicates a higher degree of correlation between sales and marketing expenditure. Therefore, the coefficient of' determination is a more objective measure of the degree of relationship.
6. The sum of r and K never adds to one, unless one of the two is zero. That is, $r + K$ can be unity either when there is no correlation or when there is perfect correlation. Except in these two extreme situations, $(r + K) > 1$.

5.7 CORRELATION ANALYSIS VERSUS REGRESSION ANALYSIS

Correlation and Regression are the two related aspects of a single problem of the analysis of the relationship between the variables. If we have information on more than one variable, we might be interested in seeing if there is any connection - any association - between them. If we found such a association, we might again be interested in predicting the value of one variable for the given and known values of other variable(s).

1. Correlation literally means the relationship between two or more variables that vary in sympathy so that the movements in one tend to be accompanied by the corresponding movements in the other(s). On the other hand, regression means stepping back or returning to the average value and is a mathematical measure expressing the average relationship between the two variables.
2. Correlation coefficient r_{xy} between two variables X and Y is a measure of the direction and degree of the linear relationship between two variables that is mutual. It is symmetric, *i.e.,* $r_{yx} = r_{xy}$ and it is immaterial which of X and Y is dependent variable and which is independent variable.

   Regression analysis aims at establishing the functional relationship between the two(or more) variables under study and then using this relationship to predict or estimate the value of the dependent variable for any given value of the independent variable(s). It also reflects upon the nature of the variable, *i.e.,* which is dependent variable and which is independent variable. Regression coefficient are not symmetric in X and Y, *i.e.,* $b_{yx} \neq b_{xy}$.
3. Correlation need not imply cause and effect relationship between the variable under study. However, regression analysis clearly indicates the cause and effect relationship between the variables. The variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.
4. Correlation coefficient r_{xy} is a relative measure of the linear relationship between X and Y and is independent of the units of measurement. It is a pure number lying between ±1.

On the other hand, the regression coefficients, b_{yx} and b_{xy} are absolute measures representing the change in the value of the variable Y (or X), for a unit change in the value of the variable X (or Y). Once the functional form of regression curve is known; by substituting the value of the independent variable we can obtain the value of the dependent variable and this value will be in the units of measurement of the dependent variable.

5. There may be non-sense correlation between two variables that is due to pure chance and has no practical relevance, *e.g.,* the correlation, between the size of shoe and the intelligence of a group of individuals. There is no such thing like non-sense regression.

5.8 SOLVED PROBLEMS

Example 5-1

The following table shows the number of motor registrations in a certain territory for a term of 5 years and the sale of motor tyres by a firm in that territory for the same period.

Year	Motor Registrations	No. of Tyres Sold
1	600	1,250
2	630	1,100
3	720	1,300
4	750	1,350
5	800	1,500

Find the regression equation to estimate the sale of tyres when the motor registration is known. Estimate sale of tyres when registration is 850.

Solution: Here the dependent variable is number of tyres; dependent on motor registrations. Hence we put motor registrations as X and sales of tyres as Y and we have to establish the regression line of Y on X.

Calculations of values for the regression equation are given below:

X	Y	$d_x = X-\overline{X}$	$d_y = Y-\overline{Y}$	d_x^2	$d_x\, d_y$
600	*1,250*	*-100*	*-50*	*10,000*	*5,000*
630	*1,100*	*-70*	*-200*	*4,900*	*14,000*
720	*1,300*	*20*	*0*	*400*	*0*
750	*1,350*	*50*	*50*	*2,500*	*2,500*
800	*1,500*	*100*	*200*	*10,000*	*20,000*
$\sum X = 3{,}500$	$\sum Y = 6{,}500$	$\sum d_x = 0$	$\sum d_y = 0$	$\sum d_x^2 = 27{,}800$	$\sum d_x d_y = 41{,}500$

$$\overline{X} = \frac{\sum X}{N} = \frac{3{,}500}{5} = 700 \qquad \text{and} \qquad \overline{Y} = \frac{\sum Y}{N} = \frac{6{,}500}{5} = 1{,}300$$

$b_{yx} =$ Regression coefficient of Y on X

$$b_{yx} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2} = \frac{\sum d_x d_y}{\sum d_x^2} = \frac{4{,}1500}{2{,}7800} = 1.4928$$

Now we can use these values for the regression line

$$Y-\overline{Y} = b_{yx}\ (X-\overline{X})$$

or $$Y - 1300 = 1.4928\ (X - 700)$$

$$Y = 1.4928\,X + 255.04$$

When $X = 850$, the value of Y can be calculated from the above equation, by putting $X = 850$ in the equation.

$$Y = 1.4928\ x\ 850 + 255.\ 04$$

$$= 1523.92$$

$$= 1{,}524 \text{ Tyres}$$

Example 5-2

A panel of Judges A and B graded seven debators and independently awarded the following marks:

Debator	Marks by A	Marks by B
1	40	32
2	34	39

3	28	26
4	30	30
5	44	38
6	38	34
7	31	28

An eighth debator was awarded 36 marks by judge A, while Judge B was not present. If Judge B were also present, how many marks would you expect him to award to the eighth debator, assuming that the same degree of relationship exists in their judgement?

Solution: Let us use marks from Judge A as *X* and those from Judge B as *Y*. Now we have to work out the regression line of *Y* on *X* from the calculation below:

Debtor	X	Y	U = X-35	V = Y-30	U^2	V^2	UV
1	*40*	*32*	*5*	*2*	*25*	*4*	*10*
2	*34*	*39*	*-1*	*9*	*1*	*81*	*-9*
3	*28*	*26*	*-7*	*-4*	*49*	*16*	*28*
4	*30*	*30*	*-5*	*0*	*25*	*0*	*0*
5	*44*	*38*	*9*	*8*	*81*	*64*	*72*
6	*38*	*34*	*3*	*4*	*9*	*16*	*12*
7	*31*	*28*	*-4*	*-2*	*16*	*4*	*8*
N = 7			$\sum U = 0$	$\sum V = 17$	$\sum U^2 = 206$	$\sum V^2 = 185$	$\sum UV = 121$

$\overline{X} = A + \dfrac{\sum U}{N} = 35 + \dfrac{0}{7} = 35$ and $\overline{Y} = A + \dfrac{\sum V}{N} = 30 + \dfrac{17}{7} = 32.43$

$$b_{yx} = b_{vu} = \frac{N\sum UV - \left(\sum U \sum V\right)}{N\sum U^2 - \left(\sum U\right)^2}$$

$$= \frac{7x121 - 0x17}{7x206 - 0} = 0.587$$

Hence regression equation can be written as

$$Y - \overline{Y} = b_{yx}\ (X - \overline{X})$$

$$Y - 32.43 = 0.587\ (X - 35)$$

$$or \quad Y = 0.587X + 11.87$$

When $X = 36$ (awarded by Judge A)

$$Y = 0.587 \, x \, 36 + 11.87$$

$$= 33$$

Thus if Judge B were present, he would have awarded 33 marks to the eighth debator.

Example 5-3

For some bivariate data, the following results were obtained.

Mean value of variable X	=	53.2
Mean value of variable Y	=	27.9
Regression coefficient of Y on X	=	- 1.5
Regression coefficient of X on Y	=	- 0.2

What is the most likely value of Y, when $X = 60$?

What is the coefficient of correlation between X and Y?

Solution: Given data indicate

$$\overline{X} = 53.2 \qquad \overline{Y} = 27.9$$

$$b_{yx} = -1.5 \qquad b_{xy} = -0.2$$

To obtain value of Y for $X = 60$, we establish the regression line of Y on X,

$$Y - \overline{Y} = b_{yx} \; (X - \overline{X})$$

$$Y - 27.9 = -1.5 \; (X - 53.2)$$

$$or \qquad Y = -1.5X + 107.7$$

Putting value of $X = 60$, we obtain

$$Y = -1.5 \, x \, 60 + 107.7$$

$$= 17.7$$

Coefficient of correlation between X and Y is given by G.M. of b_{yx} and b_{xy}

$$r^2 = b_{yx} \, b_{xy}$$

$$= (-1.5)\,x\,(-0.2)$$

$$= 0.3$$

So $\quad r \quad = \pm\sqrt{0.3} = \pm\, 0.5477$

Since both the regression coefficients are negative, we assign negative value to the correlation coefficient

$$r \quad = -\, 0.5477$$

Example 5-4

Write regression equations of *X* on *Y* and of *Y* on *X* for the following data

X:	45	48	50	55	65	70	75	72	80	85
Y:	25	30	35	30	40	50	45	55	60	65

Solution: We prepare the table for working out the values for the regression lines.

X	Y	U = X-65	V = Y-45	U^2	UV	V^2
45	*25*	*-20*	*-20*	*400*	*400*	*400*
48	*30*	*-17*	*-15*	*289*	*255*	*225*
50	*35*	*-15*	*-10*	*225*	*150*	*100*
55	*30*	*-10*	*-15*	*100*	*150*	*225*
65	*40*	*0*	*-5*	*0*	*0*	*25*
70	*50*	*5*	*5*	*25*	*25*	*25*
75	*45*	*10*	*0*	*100*	*0*	*0*
72	*55*	*7*	*5*	*49*	*35*	*25*
80	*60*	*15*	*15*	*225*	*225*	*225*
85	*65*	*20*	*20*	*400*	*400*	*400*
$\sum X = 645$	$\sum Y = 435$	$\sum U = 5$	$\sum V = -20$	$\sum U^2 = 1813$	$\sum V^2 = 1415$	$\sum UV = 1675$

We have,

$$\overline{X} = \frac{\sum X}{N} = \frac{645}{10} = 64.5 \quad \text{and} \quad \overline{Y} = \frac{\sum Y}{N} = \frac{435}{10} = 43.5$$

$$b_{yx} = \frac{N\sum UV - \left(\sum U \sum V\right)}{N\sum U^2 - \left(\sum U\right)^2}$$

$$= \frac{(10)\,x\,1415 - (5)\,x\,(-20)}{(10)\,x\,1813 - (5)^2}$$

$$= \frac{14150 + 100}{18130 - 25} = \frac{14250}{18105} = 0.787$$

Regression equation of Y on X is

$$Y - \overline{Y} = b_{yx}\ (X - \overline{X})$$

$$Y - 43.5 = 0.787\ (X - 64.5)$$

or $$Y = 0.787X + 7.26$$

Similarly b_{xy} can be calculated as

$$b_{xy} = \frac{N\sum UV - \left(\sum U \sum V\right)}{N\sum V^2 - \left(\sum V\right)^2}$$

$$= \frac{(10)\,x\,1415 - (5)\,x\,(-20)}{(10)\,x\,1675 - (-20)^2}$$

$$= \frac{14150 + 100}{16750 - 400} = \frac{14250}{16350} = 0.87$$

Regression equation of X on Y will be

$$X - \overline{X} = b_{xy}\ (\mathrm{Y} - \overline{Y})$$

$$X - 64.5 = 0.87\ (Y - 43.5)$$

or $$X = 0.87Y + 26.65$$

Example 5-5

The lines of regression of a bivariate population are

$$8X - 10Y + 66 = 0$$

$$40X - 18Y = 214$$

The variance of X is 9. Find

(i) The mean value of X and Y

(ii) Correlation coefficient between X and Y

(iii) Standard deviation of Y

Solution: The regression lines given are

$$8X - 10Y + 66 = 0$$

$$40X - 18Y \ = 214$$

Since both the lines of regression pass through the mean values, the point ($\overline{X}$,$\overline{Y}$) will satisfy both the equations.

Hence these equations can be written as

$$8\ \overline{X}\ - 10\ \overline{Y}\ + 66 = 0$$

$$40\ \overline{X}\ - 18\overline{Y}\ - 214 = 0$$

Solving these two equations for $\overline{X}$ and $\overline{Y}$, we obtain

$\overline{X} = 13$ and $\overline{Y} = 17$

(ii) For correlation coefficient between X and Y, we have to calculate the values of b_{yx} and b_{xy}

Rewriting the equations

$$10Y = 8X + 66$$

$$b_{yx} = + 8/10 = + 4/5$$

Similarly, $40X = 18Y + 214$

$$b_{xy} = 18/40 = 9/20$$

By these values, we can now work out the correlation coefficient.

$$r^2 = b_{yx} \,.\, b_{xy}$$

$$= 4/5 \text{ x } 9/20 = 9/25$$

So $r = \underline{+}\sqrt{9/25}$

$$= \underline{+}\ 0.6$$

Both the values of the regression coefficients being positive, we have to consider only the positive value of the correlation coefficient. Hence $r = 0.6$

(iii) We have been given variance of X *i.e* $S_x^{\ 2} = 9$

$S_x = \pm 3$

We consider $S_x = 3$ as SD is always positive

Since $b_{yx} = r\, S_y / S_x$

Substituting the values of b_{yx}, r and S_x we obtain,

$S_y = 4/5\ x\ 3/0.6$

$= 4$

Example 5-6

The height of a child increases at a rate given in the table below. Fit the straight line using the method of least-square and calculate the average increase and the standard error of estimate.

Month:	1	2	3	4	5	6	7	8	9	10
Height:	52.5	58.7	65	70.2	75.4	81.1	87.2	95.5	102.2	108.4

Solution: For Regression calculations, we draw the following table

Month (X)	*Height (Y)*	X^2	XY
1	*52.5*	*1*	*52.5*
2	*58.7*	*4*	*117.4*
3	*65.0*	*9*	*195.0*
4	*70.2*	*16*	*280.8*
5	*75.4*	*25*	*377.0*
6	*81.1*	*36*	*486.6*
7	*87.2*	*49*	*610.4*
8	*95.5*	*64*	*764.0*
9	*102.2*	*81*	*919.8*
10	*108.4*	*100*	*1084.0*
$\sum X = 55$	$\sum Y = 796.2$	$\sum X^2 = 385$	$\sum XY = 4887.5$

Considering the regression line as $Y = a + bX$, we can obtain the values of a and b from the above values.

$$a = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N\sum X^2 - \left(\sum X\right)^2}$$

$$= \frac{796.2\, x\, 385 \quad - 55 \quad x\, 4887.5}{10\, x\, 385 - 55\, x\, 55}$$

$$= 45.73$$

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2}$$

$$= \frac{10\, x\, 4887.5 - 55 \quad x\, 796.2}{10\, x\, 385 - 55\, x\, 55}$$

$$= 6.16$$

Hence the regression line can be written as

$$Y = 45.73 + 6.16X$$

For standard error of estimation, we note the calculated values of the variable against the observed values,

When $X = 1$, $Y_1 = 45.73 + 6.16 = 51.89$

for $X = 2$, $Y_2 = 45.73 + 616\ x\ 2 = 58.05$

Other values for $X = 3$ to $X = 10$ are calculated and are tabulated as follows:

Month (X)	*Height (Y)*	Y_i	$Y-Y_i$	$(Y-Y_i)^2$
1	*52.5*	*51.89*	*0.61*	*0.372*
2	*58.7*	*58.05*	*0.65*	*0.423*
3	*65.0*	*64.21*	*0.79*	*0.624*
4	*70.2*	*70.37*	*-0.17*	*0.029*
5	*75.4*	*76.53*	*-1.13*	*1.277*
6	*81.1*	*82.69*	*-1.59*	*2.528*
7	*87.2*	*88.85*	*-1.65*	*2.723*
8	*95.5*	*95.01*	*0.49*	*0.240*
9	*102.2*	*101.17*	*1.03*	*1.061*
10	*108.4*	*107.33*	*1.07*	*1.145*

	$\sum (Y - Y_i)^2 = 10.421$

Standard error of estimation

$$S_{yx} = \sqrt{\frac{1}{N}\sum (Y - Y_i)^2}$$

$$= \sqrt{\frac{10.421}{10}}$$

$$= 1.02$$

Example 5-7

Given $X = 4Y+5$ and $Y = \text{k}X + 4$ are the lines of regression of X on Y and of Y on X respectively. If k is positive, prove that it cannot exceed ¼.

If k = 1/16, find the means of the two variables and coefficient of correlation between them.

Solution: Line $X = 4Y + 5$ is regression line of X on Y

So $b_{xy} = 4$

Similarly from regression line of Y on X, $Y = \text{k}X + 4$,

We get $b_{yx} = \text{k}$

Now

$$r^2 = b_{xy}.\, b_{yx}$$

$$= 4\text{k}$$

Since $0 \le r^2 \le 1$, we obtain $0 \le 4k \le 1$,

Or $0 \le k \le \frac{1}{4}$,

Now for k = $\frac{1}{16}$,

$$r^2 = 4x\frac{1}{16} = \frac{1}{4}$$

$r = +$ ½

= ½ since b_{yx} and b_{yx} are positive

When k = $\frac{1}{16}$, the regression line of Y on X becomes

$$Y = \frac{1}{16}X + 4$$

Or $\qquad X - 16Y + 64 = 0$

Since line of regression pass through the mean values of the variables, we obtain revised equations as

$$\overline{X} - 4\overline{Y} - 5 = 0$$

$$\overline{X} - 16\overline{Y} + 64 = 0$$

Solving these two equations, we get

$$\overline{X} = 28 \qquad \text{and} \qquad \overline{Y} = 5.75$$

Example 5-8

A firm knows from its past experience that its monthly average expenses (X) on advertisement are Rs 25,000 with standard deviation of Rs 25.25. Similarly, its average monthly product sales (Y) have been Rs 45,000 with standard deviation of Rs 50.50. Given this information and also the coefficient of correlation between sales and advertisement expenditure as 0.75, estimate

(i) the most appropriate value of sales against an advertisement expenditure of Rs 50,000

(ii) the most appropriate advertisement expenditure for achieving a sales target of Rs 80,000

Solution: Given the following

$\overline{X}$ = Rs 25,000 $\qquad S_x$ = Rs 25.25

$\overline{Y}$ = Rs 45,000 $\qquad S_y$ = Rs 50.50

$r = 0.75$

(i) Using equation $Y_c - \overline{Y} = r \frac{S_y}{S_x} (X - \overline{X})$, the most appropriate value of sales Y_c for an advertisement expenditure X = Rs 50,000 is

$$Y_c - 45{,}000 = 0.75 \frac{50.50}{25.25}(50{,}000 - 25{,}000)$$

$$Y_c = 45{,}000 + 37{,}500$$

$$= \text{Rs } 82{,}500$$

(ii) Using equation $X_c - \overline{X} = r \frac{S_x}{S_y}(Y - \overline{Y})$, the most appropriate value of advertisement expenditure X_c for achieving a sales target Y= Rs 80,000 is

$$X_c - 25{,}000 = 0.75 \frac{25.25}{50.50}(80{,}000 - 45{,}000)$$

$$X_c = 13{,}125 + 25{,}000$$

$$= \text{Rs } 38{,}125$$

Objectives : **The overall objective of this lesson is to give an understanding of Index Numbers. After successful completion of the lesson, the students will be able to understand the concepts, techniques and the problems involved in constructing and using index numbers.**

Structure

6.1 INTRODUCTION

In business, managers and other decision makers may be concerned with the way in which the values of variables change over time like prices paid for raw materials, numbers of employees and customers, annual income and profits, and so on. Index numbers are one way of describing such changes.

If we turn to any journal devoted to economic and financial matters, we are very likely to come across an index number of one or the other type. It may be an index number of share prices or a wholesale price index or a consumer price index or an index of industrial production. The objective of these index numbers is to measure the changes that have occurred in prices, cost of living, production, and so forth. *For example,* if a wholesale price index number for the year 2000 with base year 1990 was 170; it shows that wholesale prices, in general, increased by 70 percent in 2000 as compared to those in 1990. Now, if the same index number moves to 180 in 2001, it shows that there has been 80 percent increase in wholesale prices in 2001 as compared to those in 1990.

With the help of various index numbers, economists and businessmen are able to describe and appreciate business and economic situations quantitatively. Index numbers were originally developed by economists for monitoring and comparing different groups of goods. It is necessary in business to understand and manipulate the different published index serieses, and to construct index series of your own. Having constructed your own index, it can then be compared to a national one such as the RPI, a similar index for your industry as a whole and also to indexes for your competitors. These comparisons are a very useful tool for decision making in business.

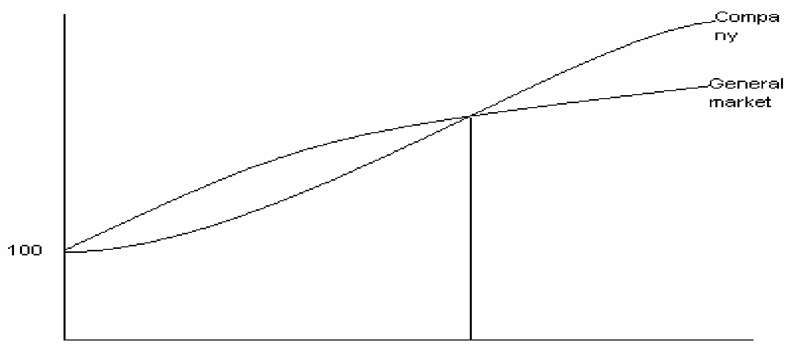


Figure 6-1 The Indexes of the Volume of Sales

For example, an accountant of a supermarket chain could construct an index of the company's own sales and compare it to the index of the volume of sales for the general supermarket industry. A graph of the two indexes will illustrate the company's performance within the sector. It is immediately clear from Figure 6-1 that, after initially lagging behind the general market, the supermarket company caught up and then overtook it. In the later stages, the company was having better results than the general market but that, as with the whole industry, those had levelled out.

Our focus in this lesson will be on the discussion related to the methodology of index number construction. The scope of the lesson is rather limited and as such, it does not discuss a large number of index numbers that are presently compiled and published by different departments of the Government of India.

6.2 WHAT ARE INDEX NUMBERS?

"Index numbers are statistical devices designed to measure the relative changes in the level of a certain phenomenon in two or more situations". The phenomenon under consideration may be any field of quantitative measurements. It may refer to a single variable or a group of distinct but related variables. In Business and Economics, the phenomenon under consideration may be:

- ✓ the prices of a particular commodity like steel, gold, leather, *etc.* or a group of commodities like consumer goods, cereals, milk and milk products, cosmetics, *etc.*
- ✓ volume of trade, factory production, industrial or agricultural production, imports or exports, stocks and shares, sales and profits of a business house and so on.

- ✓ the national income of a country, wage structure of workers in various sectors, bank deposits, foreign exchange reserves, cost of living of persons of a particular community, class or profession and so on.

The various situations requiring comparison may refer to either

- ✓ the changes occurring over a time, or
- ✓ the difference(s) between two or more places, or
- ✓ the variations between similar categories of objects/subjects, such as persons, groups of persons, organisations *etc.* or other characteristics such as income, profession, *etc.*

The utility of index numbers in facilitating comparison may be seen when, *for example* we are interested in studying the general change in the price level of consumer goods, *i.e.* good or commodities consumed by the people belonging to a particular section of society, say, low income group or middle income group or labour class and so on. Obviously these changes are not directly measurable as the price quotations of the various commodities are available in different units, *e.g.,* cereals (wheat, rice, pulses, *etc*) are quoted in Rs per quintal or kg; water in Rs per gallon; milk, petrol, kerosene, *etc.* in Rs per liter; cloth in Rs per miter and so on.

Further, the prices of some of the commodities may be increasing while those of others may be decreasing during the two periods and the rates of increase or decrease may be different for different commodities. Index number is a statistical device, which enables us to arrive at a single representative figure that gives the general level of the price of the phenomenon (commodities) in an extensive group. According to Wheldon:

> *"Index number is a statistical device for indicating the relative movements of the data where measurement of actual movements is difficult or incapable of being made."*

FY Edgeworth gave the classical definition of index numbers as follows:

> *"Index number shows by its variations the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice."*

On the basis of above discussion, the following characteristics of index numbers are apparent:

1. ***Index Numbers are specialized averages:*** An average is a summary figure measuring the central tendency of the data, representing a group of figures. Index number has all these functions to perform. L R Connor states, *"in its simplest form, it (index number) represents a special case of an average, generally a weighted average compiled from a sample of items judged to be representative of the whole".* It is a special type of average – it averages variables having different units of measurement.

2. ***Index Numbers are expressed in percentages:*** Index numbers are expressed in terms of percentages so as to show the extent of change. However, percentage sign (%) is never used.

3. ***Index Numbers measure changes not capable of direct measurement****:* The technique of index numbers is utilized in measuring changes in magnitude, which are not capable of direct measurement. Such magnitudes do not exist in themselves. Examples of such magnitudes are 'price level', 'cost of living', 'business or economic activity' *etc.* The statistical methods used in the construction of index numbers are largely methods for combining a number of phenomena representing a particular magnitude in such a manner that the changes in that magnitude may be measured in a meaningful way without introduction of serious bias.

4. ***Index Numbers are for comparison:*** The index numbers by their nature are comparative. They compare changes taking place over time or between places or between like categories.

In brief, index number is a statistical technique used in measuring the composite change in several similar economic variables over time. It measures only the composite change, because some of the variables included may be showing an increase, while some others may be showing a decrease. It synthesizes the changes taking place in different directions and by varying extents into the one composite change. Thus, an index number is a device to simplify comparison to show relative movements of the data concerned and to replace what may be complicated figures by simple ones calculated on a percentage basis.

6.3 USES OF INDEX NUMBER

The first index number was constructed by an Italian, Mr G R Carli, in 1764 to compare the changes in price for the year 1750 (current year) with the price level in 1500 (base year) in order to study the effect of discovery of America on the price level in Italy. Though originally designed to study the general level of prices or accordingly purchasing power of money, today index numbers are extensively used for a variety of purposes in economics, business, management, *etc.,* and for quantitative data relating to production, consumption, profits, personnel and financial matters *etc.,* for comparing changes in the level of phenomenon for two periods, places, *etc.* In fact there is hardly any field or quantitative measurements where index numbers are not constructed. They are used in almost all sciences – natural, social and physical. The main uses of index numbers can be summarized as follows:

1. **Index Numbers as Economic Barometers**

   Index numbers are indispensable tools for the management personnel of any government organisation or individual business concern and in business planning and formulation of executive decisions. The indices of prices (wholesale & retail), output (volume of trade, import and export, industrial and agricultural production) and bank deposits, foreign exchange and reserves *etc.,* throw light on the nature of, and variation in the general economic and business activity of the country. They are the indicators of business environment. A careful study of these indices gives us a fairly good appraisal of the general trade, economic development and business activity of the country. In the world of G Simpson and F Kafka:

> "Index numbers are today one of the most widely used statistical devices. They are used to take the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies."

Like barometers, which are used in Physics and Chemistry to measure atmospheric pressure, index numbers are rightly termed as "economic barometers", which measure the pressure of economic and business behaviour.

2. Index Numbers Help in Studying Trends and Tendencies

Since the index numbers study the relative change in the level of a phenomenon at different periods of time, they are especially useful for the study of the general trend for a group phenomenon in time series data. The indices of output (industrial and agricultural production), volume of trade, import and export, *etc.,* are extremely useful for studying the changes in the level of phenomenon due to the various components of a time series, *viz.* secular trend, seasonal and cyclical variations and irregular components and reflect upon the general trend of production and business activity. As a measure of average change in extensive group, the index numbers can be used to forecast future events. For instance, if a businessman is interested in establishing a new undertaking, the study of the trend of changes in the prices, wages and incomes in different industries is extremely helpful to him to frame a general idea of the comparative courses, which the future holds for different undertakings.

3. Index Numbers Help in Formulating Decisions and Policies

Index numbers of the data relating to various business and economic variables serve an important guide to the formulation of appropriate policy. *For example,* the cost of living index numbers are used by the government and, the industrial and business concerns for the regulation of dearness allowance (D.A.) or grant of bonus to the workers so as to enable them to meet the increased cost of living from time to time. The excise duty on the production or sales of a commodity is regulated according to the index numbers of the consumption of the commodity from time to time. Similarly, the indices of consumption of various commodities help in the planning of their future production. Although index numbers are now widely used to study the general economic and business conditions of the society, they are also applied with advantage by sociologists (population indices), psychologists (IQs'), health and educational authorities *etc.,* for formulating and revising their policies from time to time.

4. **Price Indices Measure the Purchasing Power of Money**

A traditional use of index numbers is in measuring the purchasing power of money. Since the changes in prices and purchasing power of money are inversely related, an increase in the general price index indicates that the purchasing power of money has gone down.

In general, the purchasing power of money may be computed as

$$\text{Purchasing Power} = \frac{1}{\text{General Price Index}} x100$$

Accordingly, if the consumer price index for a given year is 150, the purchasing power of a rupee is (1/150) 100 = 0.67. That is, the purchasing power of a rupee in the given year is 67 paise as compared to the base year.

With the increase in prices, the amount of goods and services which money wages can buy (or the real wages) goes on decreasing. Index numbers tell us the change in real wages, which are obtained as

$$\text{Real Wage} = \frac{\text{Money Wage}}{\text{Consumer Price Index}} x100$$

A real wage index equal to, say, 120 corresponding to money wage index of 160 will indicate an increase in real wages by only 20 per cent as against 60 per cent increase in money wages.

Index numbers also serve as the basis of determining the terms of exchange. The terms of exchange are the parity rate at which one set of commodities is exchanged for another set of commodities. It is determined by taking the ratio of the price index for the two groups of commodities and expressing it in percentage.

For example, if A and B are the two groups of commodities with 120 and 150 as their price index in a particular year, respectively, the ratio 120/150 multiplied by 100 is 80 per cent. It means that prices of A group of commodities in terms of those in group B are lower by 20 per cent.

5. Index Numbers are Used for Deflation

Consumer price indices or cost of living index numbers are used for deflation of net national product, income value series in national accounts. The technique of obtaining real wages from the given nominal wages (as explained in use 4 above) can be used to find real income from inflated money income, real sales from nominal sales and so on by taking into account appropriate index numbers.

5.4 TYPES OF INDEX NUMBERS

Index numbers may be broadly classified into various categories depending upon the type of the phenomenon they study. Although index numbers can be constructed for measuring relative changes in any field of quantitative measurement, we shall primarily confine the discussion to the data relating to economics and business *i.e.,* data relating to prices, production (output) and consumption. In this context index numbers may be broadly classified into the following three categories:

1. **Price Index Numbers:** The price index numbers measure the general changes in the prices. They are further sub-divided into the following classes:

   (i) Wholesale Price Index Numbers: The wholesale price index numbers reflect the changes in the general price level of a country.

   (ii) Retail Price Index Numbers: These indices reflect the general changes in the retail prices of various commodities such as consumption goods, stocks and shares, bank deposits, government bonds, *etc.*

   (iii) Consumer Price Index: Commonly known as the Cost of living Index, CPI is a specialized kind of retail price index and enables us to study the effect of changes in the price of a basket of goods or commodities on the purchasing power or cost of living of a particular class or section of the people like labour class, industrial or agricultural worker, low income or middle income class *etc.*

2. **Quantity Index Numbers**: Quantity index numbers study the changes in the volume of goods produced (manufactured), consumed or distributed, like: the indices of agricultural production, industrial production, imports and exports, *etc.* They are extremely helpful in studying the level of physical output in an economy.

3. **Value Index Numbers**: These are intended to study the change in the total value (price multiplied by quantity) of output such as indices of retail sales or profits or inventories. However, these indices are not as common as price and quantity indices.

Notations Used

Since index numbers are computed for prices, quantities, and values, these are denoted by the lower case letters:

p, *q*, and *v* represent respectively the price, the quantity, and the value of an individual commodity.

Subscripts *0, 1, 2,... i, ...* are attached to these lower case letters to distinguish price, quantity, or value in any one period from those in the other. Thus,

p_0 denotes the price of a commodity in the base period,

p_1 denotes the price of a commodity in period 1, or the current period, and

p_i denotes the price of a commodity in the i^{th} period, where $i = 1,2,3, ...$

Similar meanings are assigned to $q_0, q_1, ... q_i, ...$ and $v_0, v_1, ... v_i, ...$

Capital letters *P, Q and V* are used to represent the price, quantity, and value index numbers, respectively. Subscripts attached to *P, Q,* and *V* indicates the years compared. Thus,

P_{01} means the price index for period 1 relative to period 0,

P_{02} means the price index for period 2 relative to period 0,

P_{12} means the price index for period 2 relative to period 1, and so on.

Similar meanings are assigned to quantity *Q* and value *V* indices. It may be noted that all indices are expressed in percent with 100 as the index for the base period, the period with which comparison is to be made.

Various indices can also be distinguished on the basis of the number of commodities that go into the construction of an index. Indices constructed for individual commodities or variable are termed as ***simple index numbers***. Those constructed for a group of commodities or variables are known as ***aggregative (or composite) index numbers***.

Here, in this lesson, we will develop methods of constructing simple as well as composite indices.

6.5 SIMPLE INDEX NUMBERS

A simple price index number is based on the price or quantity of a single commodity. To construct a simple index, we first have to decide on the base period and then find ratio of the value at any subsequent period to the value in that base period - *the price/quantity relative.* This ratio is then finally converted to a percentage

$$\text{Index for any Period } i = \frac{\text{Value in Period } i}{\text{Value in Base Year}} x100$$

i.e. Simple Price Index for period $i = 1,2,3$ ... will be

$$P_{0i} = \frac{p_i}{p_0} x100 \qquad \ldots\ldots\ldots\ldots(6\text{-}1)$$

Similarly, Simple Quantity Index for period $i = 1,2,3$ ... will be

$$Q_{0i} = \frac{q_i}{q_0} x100 \qquad \ldots\ldots\ldots\ldots(6\text{-}2)$$

Example 6-1

Given are the following price-quantity data of fish, with price quoted in Rs per kg and production in qtls.

Year	:	1980	1981	1982	1983	1984	1985
Price	:	15	17	16	18	22	20
Production	:	500	550	480	610	650	600

Construct:

(a) the price index for each year taking price of 1980 as base,

(b) the quantity index for each year taking quantity of 1980 as base.

Solution:

Simple Price and Quantity Indices of Fish
(Base Year = 1980)

Year	Price *(p_i)*	Quantity *(q_i)*	Price Index $P_{0i} = \frac{p_i}{p_0} x100$	Quantity Index $Q_{0i} = \frac{q_i}{q_0} x100$

1980	15	500	100.00	100.00
1981	17	550	113.33	110.00
1982	16	480	106.66	96.00
1983	18	610	120.00	122.00
1984	22	650	146.66	130.00
1985	20	600	133.33	120.00

These simple indices facilitate comparison by transforming absolute quantities/prices into percentages. Given such an index, it is easy to find the percent by which the price/quantity may have changed in a given period as compared to the base period. *For example,* observing the index computed in Example 6-1, one can firmly say that the output of fish was 30 per cent more in 1984 as compared to 1980.

It may also be noted that given the simple price/quantity for the base year and the index for the period *i = 1, 2, 3, ...;* the actual price/quantity for the period *i = 1, 2, 3, ...* may easily be obtained as:

$$p_i = p_0\left(\frac{P_{0i}}{100}\right) \qquad \text{…………}(6\text{-}3)$$

and

$$q_i = q_0\left(\frac{Q_{0i}}{100}\right) \qquad \text{…………}(6\text{-}4)$$

For example, with $i = 1983$, $Q_{0i} = 122.00$, and $q_0 = 500$,

$$q_i = 500\left(\frac{122.00}{100}\right)$$
$$= 610$$

6.6 COMPOSITE INDEX NUMBERS

The preceding discussion was confined to only one commodity. What about price/quantity changes in several commodities? In such cases, composite index numbers are used. Depending upon the method used for constructing an index, composite indices may be:

1. Simple Aggregative Price/ Quantity Index
2. Index of Average of Price/Quantity Relatives
3. Weighted Aggregative Price/ Quantity Index
4. Index of Weighted Average of Price/Quantity Relatives

6.6.1 SIMPLE AGGREGATIVE PRICE/ QUANTITY INDEX

Irrespective of the units in which prices/quantities are quoted, this index for given prices/quantities, of a group of commodities is constructed in the following three steps:

(i) *Find the aggregate of prices/quantities of all commodities for each period (or place).*

(ii) Selecting one period as the base, divide the aggregate prices/quantities corresponding to each period (or place) by the aggregate of prices/ quantities in the base period.

(iii) Express the result in percent by multiplying by 100.

The computation procedure contained in the above steps can be expressed as:

$$P_{0i} = \frac{\sum p_i}{\sum p_0} x100 \qquad \text{…………}(6\text{-}5)$$

and

$$Q_{0i} = \frac{\sum q_i}{\sum q_0} x100 \qquad \text{…………}(6\text{-}6)$$

Example 6-2

Given are the following price-quantity data, with price quoted in Rs per kg and production in qtls.

	1980		1985	
Item	Price	Production	Price	Production
Fish	15	500	20	600
Mutton	18	590	23	640

Chicken	22	450	24	500

Find (a) Simple Aggregative Price Index with 1980 as the base.

(b) Simple Aggregative Quantity Index with 1980 as the base.

Solution:

Calculations for
Simple Aggregative Price and Quantity Indices
(Base Year = 1980)

Item	Prices		*Quantities*	
	1980*(p_0)*	1985*(p_i)*	1980*(q_0)*	1985*(q_i)*
Fish	15	20	500	600
Mutton	18	23	590	640
Chicken	22	24	450	500
Sum →	55	67	1540	1740

(a) Simple Aggregative Price Index with 1980 as the base

$$P_{0i} = \frac{\sum p_i}{\sum p_0} x100$$

$$P_{0i} = \frac{67}{55} x100$$

$$P_{0i} = 121.82$$

(b) Simple Aggregative Quantity Index with 1980 as the base

$$Q_{0i} = \frac{\sum q_i}{\sum q_0} x100$$

$$Q_{0i} = \frac{1740}{1540} x100$$

$$Q_{0i} = 112.98$$

Although Simple Aggregative Index is simple to calculate, it has two important limitations:

First, equal weights get assigned to every item entering into the construction of this index irrespective of relative importance of each individual item being different. *For example,* items like pencil and milk are assigned equal importance in the construction of this index. This limitation renders the index of no practical utility.

Second, different units in which the prices are quoted also sometimes unduly affect this index. Prices quoted in higher weights, such as price of wheat per bushel as compared to a price per kg, will have unduly large influence on this index. Consequently, the prices of only a few commodities may dominate the index. This problem no longer exists when the units in which the prices of various commodities are quoted have a common base.

Even the condition of common base will provide no real solution because commodities with relatively high prices such as gold, which is not as important as milk, will continue to dominate this index excessively. *For example,* in the Example 6-2 given above chicken prices are relatively higher than those of fish, and hence chicken prices tend to influence this index relatively more than the prices of fish.

6.6.2 INDEX OF AVERAGE OF PRICE/QUANTITY RELATIVES

This index makes an improvement over the index of simple aggregative prices/quantities as it is not affected by the difference in the units of measurement in which prices/quantities are expressed. However, this also suffers from the problem of equal importance getting assigned to all the commodities.

Given the prices/quantities of a number of commodities that enter into the construction of this index, it is computed in the following two steps:

(i) After selecting the base year, find the price relative/quantity relative of each commodity for each year with respect to the base year price/quantity. As defined earlier, the price relative/quantity relative of a commodity for a given period is the ratio of the price/quantity of that commodity in the given period to its price/quantity in the base period.

(ii) Multiply the result for each commodity by 100, to get simple price/quantity indices for each commodity.

(iii) Take the average of the simple price/quantity indices by using arithmetic mean, geometric mean or median.

Thus it is computed as:

$$P_{0i} = \text{Average of}\left(\frac{p_i}{p_0}x100\right)$$

and $$Q_{0i} = \text{Average of}\left(\frac{q_i}{q_0}x100\right)$$

Using arithmetic mean

$$P_{0i} = \frac{\sum\left(\frac{p_i}{p_0}x100\right)}{N} \qquad \text{…………}\textit{(6-7)}$$

and $$Q_{0i} = \frac{\sum\left(\frac{q_i}{q_0}x100\right)}{N} \qquad \text{…………}\textit{(6-8)}$$

Using geometric mean

$$P_{0i} = Anti\log\left[\frac{1}{N}\sum\log\left(\frac{p_i}{p_0}x100\right)\right] \qquad \text{…………}\textit{(6-9)}$$

and $$Q_{0i} = Anti\log\left[\frac{1}{N}\sum\log\left(\frac{q_i}{q_0}x100\right)\right] \qquad \text{…………}\textit{(6-10)}$$

Example 6-3

From the data in Example 6.2 find:

(a) Index of Average of Price Relatives (base year 1980); using mean, median and geometric mean.

(b) Index of Average of Quantity Relatives (base year 1980); using mean, median and geometric mean.

Solution:

Calculations for
Index of Average of Price Relatives and Quantity Relatives
(Base Year = 1980)

Item	Price Relative $=\left(\frac{p_i}{p_0}x100\right)$	$\log\left(\frac{p_i}{p_0}x100\right)$	Quantity Relative $=\left(\frac{q_i}{q_0}x100\right)$	$\log\left(\frac{q_i}{q_0}x100\right)$
Fish	133.33	2.1248	120.00	2.0792
Mutton	127.77	2.1063	108.47	2.0354
Chicken	109.09	2.0378	111.11	2.0457
Sum →	370.19	6.2689	339.58	6.1603

(a) Index of Average of Price Relatives (base year 1980)

Using arithmetic mean

$$P_{0i} = \frac{\sum\left(\frac{p_i}{p_0}x100\right)}{N}$$

$$= \frac{370.19}{3}$$

$$= 123.39$$

Using Median

$$P_{0i} = \text{Size of}\left(\frac{\text{N}+1}{2}\right)th \text{ item}$$

$$= \text{Size of}\left(\frac{3+1}{2}\right)th \text{ item}$$

$$= \text{Size of } 2nd \text{ item}$$

$$= 127.77$$

Using geometric mean

$$P_{0i} = Anti\log\left[\frac{1}{N}\sum\log\left(\frac{p_i}{p_0}x100\right)\right]$$

$$= Anti\log\left[\frac{1}{3}(6.2689)\right]$$

$$= Anti\log[2.08963]$$

$$= 122.92$$

(b) Index of Average of Quantity Relatives (base year 1980)

Using arithmetic mean

$$Q_{0i} = \frac{\sum\left(\frac{q_i}{q_0} x100\right)}{N}$$

$$= \frac{339.58}{3}$$

$$= 113.19$$

Using Median

$$Q_{0i} = \text{Size of}\left(\frac{\text{N}+1}{2}\right)th \text{ item}$$

$$= \text{Size of}\left(\frac{3+1}{2}\right)th \text{ item}$$

$$= \text{Size of } 2nd \text{ item}$$

$$= 111.11$$

Using geometric mean

$$Q_{0i} = Anti\log\left[\frac{1}{N}\sum\log\left(\frac{q_i}{q_0} x100\right)\right]$$

$$= Anti\log\left[\frac{1}{3}(6.1603)\right]$$

$$= Anti\log[2.05343]$$

$$= 113.09$$

Apart from the inherent drawback that this index accords equal importance to all items entering into its construction, a simple arithmetic mean and median are not appropriate average to be applied to ratios. Because it is generally believed that a simple average injects an upward bias in the index. So geometric mean is considered a more appropriate average for ratios and percentages.

6.6.3 WEIGHTED AGGREGATIVE PRICE/QUANTITY INDICES

We have noted that the simple aggregative price/quantity indices do not take care of the differences in the weights to be assigned to different commodities that enter into their construction. It is primarily because of this limitation that the simple aggregative indices are of very limited use. Weighted aggregative Indices make up this deficiency by assigning proper weights to individual items.

Among several ways of assigning weights, two widely used ways are:

(i) to use base period quantities/prices as weights, popularly known as ***Laspeyre's Index***, and

(ii) to use the given (current) period quantities/prices as weights, popularly known as ***Paasche's Index***.

6.6.3.1 Laspeyre's Index

Laspeyre's Price Index, using base period quantities as weights is obtained as

$$P_{0i}^{La} = \frac{\sum p_i q_0}{\sum p_0 q_0} x100 \qquad \text{…………}\textit{(6-11)}$$

Laspeyre's Quantity Index, using base period prices as weights is obtained as

$$Q_{0i}^{La} = \frac{\sum q_i p_0}{\sum q_0 p_0} x100 \qquad \text{…………}\textit{(6-12)}$$

6.6.3.2 Paasche's Index

Paasche's Price Index, using base period quantities as weights is obtained as

$$P_{0i}^{Pa} = \frac{\sum p_i q_i}{\sum p_0 q_i} x100 \qquad \text{…………}\textit{(6-13)}$$

Paasche's Quantity Index, using base period prices as weights is obtained as

$$Q_{0i}^{Pa} = \frac{\sum q_i p_i}{\sum q_0 p_i} x100 \qquad \text{…………}\textit{(6-14)}$$

Example 6-4

From the data in Example 6.2 find:

(a) Laspeyre's Price Index for 1985, using 1980 as the base

(b) Laspeyre's Quantity Index for 1985, using 1980 as the base

(c) Paasche's Price Index for 1985, using 1980 as the base

(d) Paasche's Quantity Index for 1985, using 1980 as the base

Solution:

Calculations for
Laspeyre's and Paasche's Indices
(Base Year = 1980)

Item	$p_0 q_0$	$p_1 q_0$	$p_0 q_1$	$p_1 q_1$
Fish	7500	10000	9000	12000
Mutton	10620	13570	11520	14720
Chicken	9900	10800	11000	12000
Sum →	28020	34370	31520	38720

(a) Laspeyre's Price Index for 1985, using 1980 as the base

$$P_{0i}^{La} = \frac{\sum p_i q_0}{\sum p_0 q_0} x100$$

$$= \frac{34370}{28020} x100$$

$$= 122.66$$

(b) Laspeyre's Quantity Index for 1985, using 1980 as the base

$$Q_{0i}^{La} = \frac{\sum q_i p_0}{\sum q_0 p_0} x100$$

$$= \frac{31520}{28020} x100$$

$$= 112.49$$

(c) Paasche's Price Index for 1985, using 1980 as the base

$$P_{0i}^{Pa} = \frac{\sum p_i q_i}{\sum p_0 q_i} x100$$

$$= \frac{38720}{31520} x100$$

$$= 122.84$$

(d) Paasche's Quantity Index for 1985, using 1980 as the base

$$Q_{0i}^{Pa} = \frac{\sum q_i p_i}{\sum q_0 p_i} x100$$

$$= \frac{38720}{34370} x100$$

$$= 112.66$$

Interpretations of Laspeyre's Index

On close examination it will be clear that the Laspeyre's Price Index offers the following precise interpretations:

1. It compares the cost of collection of a fixed basket of goods selected in the base period with the cost of collecting the same basket of goods in the given (current) period.

   Accordingly, the cost of collection of 500 qtls of fish, 590 qtls of mutton and 450 qtls of chicken has increased by 22.66 per cent in 1985 as compared to what it was in 1980. Viewed differently, it indicates that a fixed amount of goods sold at 1985 prices yield 22.66 per cent more revenue than what it did at 1980 prices.

2. It also implies that a fixed amount of goods when purchased at 1985 prices would cost 22.66 per cent more than what it did at 1980 prices. In this interpretation, the Laspeyre's Price Index serves as the basis of constructing the cost of living index, for it tells how much more does it cost to maintain the base period standard of living at the current period prices.

Laspeyre's Quantity Index, too, has precise interpretations. It reveals the percentage change in total expenditure in the given (current) period as compared to the base period if varying amounts of the same basket of goods are sold at the base period prices. When viewed in this manner, we will be required to spend 12.49 per cent more in 1985 as compared to 1980 if the quantities of fish, mutton and chicken for 1965 are sold at the base period (1980) prices.

Interpretations of Paasche's Index

A careful examination of the Paasche's Price Index will show that this too is amenable to the following precise interpretations:

1. It compares the cost of collection of a fixed basket of goods selected in the given period with the cost of collection of the same basket of goods in the base period.

   Accordingly, the cost of collection of a fixed basket of goods containing 600 qtls of fish, 640 qtls of mutton and 500 qtls of chicken in 1985 is about 22.84 per cent more than the cost of collecting the same basket of goods in 1980. Viewed a little differently, it indicates that a fixed basket of goods sold at 1985 prices yields 22.84 per cent more revenue than what it would have earned had it been sold at the base period (1980) prices.

2. It also tells that a fixed amount of goods purchased at 1985 prices will cost 22.84 per cent more than what it would have cost if this fixed amount of goods had been sold at base period (1980) prices.

Analogously, Paasche's Quantity Index, too, has its own precise meaning. It tells the per cent change in total expenditure in the given period as compared to the base period if varying amounts of the same basket of goods are to be sold at given period prices. When so viewed, we will be required to spend 12.66 per cent more in 1985 as compared to 1980 if the quantities of fish, mutton and chicken for 1980 are sold at the given period (1985) prices.

Relationship Between Laspeyre's and Paasche's Indices

In order to understand the relationship between Laspeyre's and Paasche's Indices, the assumptions on which the two indices are based be borne in mind:

Laspeyre's index is based on the assumption that *unless there is a change in tastes and preferences, people continue to buy a fixed basket of goods irrespective of how high or low* *the prices are likely to be in the future.* Paasche's index, on the other hand, assumes that *people would have bought the same amount of a given basket of goods in the past irrespective of how high or low were the past prices.*

However, the basic contention implied in the assumptions on which the two indices are based is not true. For, people do make shifts in their purchase pattern and preferences by buying more of goods that tend to become cheaper and less of those that tend to become costlier. In view of this, the following two situations that are likely to emerge need consideration:

1. When the prices of goods that enter into the construction of these indices show a general tendency to rise, those whose prices increase more than the average increase in prices will have smaller quantities in the given period than the corresponding quantities in the base period. That is, q_i*'s* will be smaller than q_0*'s* when prices in general are rising. Consequently, Paasche's index will have relatively smaller weights than those in the Laspeyre's index and, therefore, the former (P_{0i}^{Pa}) will be smaller than the latter (P_{0i}^{La}). In other words, Paasche's index will show a relatively smaller increase when the prices in general tend to rise.
2. On the contrary, when prices in general are falling, goods whose prices show a relatively smaller fall than the average fall in prices, will have smaller quantities in the given period than the corresponding quantities in the base period. This means that q_i*'s* will be smaller than q_0*'s* when prices in general are falling. Consequently, Paasche's index will have smaller weights than those in the Laspeyre's index and, therefore, the former (P_{0i}^{Pa}) will be smaller than the latter (P_{0i}^{La}). In other words, Paasche's index will show a relatively greater fall when the prices in general tend to fall.

An important inference based on the above discussion is that ***the Paasche's index has a downward bias and the Laspeyre's index an upward bias.*** This directly follows from the fact that the Paasche's index, relative to the Laspeyre's index, shows a smaller rise when the prices in general are rising, and a greater fall when the prices in general are falling.

It may, however, be noted that when the quantity demanded increases because of change in real income, tastes and preferences, advertising, *etc.,* the prices remaining unchanged, the Paasche's index will show a higher value than the Laspeyre's index. In such situations, the Paasche's index will overstate, and the Laspeyre's will understate, the changes in prices. The former now represents the upper limit, and the latter the lower limit, of the range of price changes.

The relationship between the two indices can be derived more precisely by making use of the coefficient of linear correlation computed as:

$$r_{xy} = \frac{\frac{\sum fXY}{N} - \left(\frac{\sum fX}{N}\right)\left(\frac{\sum fY}{N}\right)}{S_x S_y} \qquad \ldots\ldots\ldots\ldots(6.15)$$

in which *X* and *Y* denote the relative price movements($\frac{p_i}{p_0}$) and relative quantity movements($\frac{q_i}{q_0}$) respectively. S_x and S_y are the standard deviations of price and quantity movements, respectively. While r_{xy} represents the coefficient of correlation between the relative price and quantity movements; *f* represents the weights assigned, that is, $p_0\, q_0$. *N* is the sum of frequencies *i. e.* $N = \sum p_0 q_0$.

Substituting the values of *X*, *Y, f* and *N* in *Eq. (6-15),* and then rearranging the expression, we have

$$r_{xy} S_x S_y = \frac{\sum p_i q_i}{\sum p_0 q_0} - \frac{\sum p_i q_0}{\sum p_0 q_0} x \frac{\sum p_0 q_i}{\sum p_0 q_0}$$

If $\frac{\sum p_i q_i}{\sum p_0 q_0} = V_{0i}$, is the index of value expanded between the base period and the i^{th} period,

then dividing both sides by $\frac{\sum p_i q_i}{\sum p_0 q_0}$ or V_{0i}, we get

$$\frac{r_{xy} S_x S_y}{V_{0i}} = 1 - \frac{\sum p_i q_0}{\sum p_0 q_0} x \frac{\sum p_0 q_i}{\sum p_i q_i}$$

$$\frac{r_{xy} S_x S_y}{V_{0i}} = 1 - P_{0i}^{La} x \frac{1}{P_{0i}^{Pa}}$$

$$\frac{P_{0i}^{La}}{P_{0i}^{Pa}} = 1 - \frac{r_{xy} S_x S_y}{V_{0i}} \qquad \ldots\ldots\ldots\ldots(6.16)$$

The relationship in *Eq. (6.16)* offers the following useful results:

1. $P_{0i}^{La} = P_{0i}^{Pa}$ when either r_{xy} , S_x and S_y is equal to zero. That is, the two indices will give the same result either when there is no correlation between the price and quantity movements, or when the price or quantity movements are in the same ratio so that S_x or S_y is equal to zero.

2. Since in actual practice r_{xy} will have a negative value between 0 and -1, and as neither $S_x = 0$ nor $S_y = 0$, the right hand side of *Eq. (6-16)* will be less than 1. This means that P_{0i}^{La} is normally greater than P_{0i}^{Pa} .

3. Given the overall movement in the index of value (V_{0i}) expanded, the greater the coefficient of correlation (r_{xy}) between price and quantity movements and/or the greater the degree of dispersion (S_x and S_y) in the price and quantity movements, the greater the discrepancy between P_{0i}^{La} and P_{0i}^{Pa} .

4. The longer the time interval between the two periods to be compared, the more the chances for price and quantity movements leading to higher values of S_x and S_y. The assumption of tastes, habits, and preferences remaining unchanged breaking down over a longer period, people do find enough time to make shifts in their consumption pattern, buying more of goods that may have become relatively cheaper and less of those that may have become relatively dearer. All this will end up with a higher degree of correlation between the price and quantity movement. Consequently, P_{0i}^{La} will diverge from P_{0i}^{Pa} more in the long run than in the short run. So long as the periods to be compared are not much apart, P_{0i}^{La} will be quite close to P_{0i}^{Pa}.

Laspeyre's and Paasche's Indices Further Considered

The use of different system of weights in these two indices may give an impression as if they are opposite to each other. Such an impression is not sound because both serve the same purpose, although they may give different results when applied to the same data.

This raises an important question. Which one of them gives more accurate results and which one should be preferred over the other? The answer to this question is rather difficult since both the indices are amenable to precise and useful results.

Despite a very useful and precise difference in interpretation, in actual practice the Laspeyre's index is used more frequently than the Paasche's index for the simple reason that the latter requires frequent revision to take into account the yearly changes in weights. No such revision is required in the case of the Laspeyre's index where once the weights have been determined, these do not require any change in any subsequent period. It is on this count that the Laspeyre's index is preferred over the Paasche's index.

However, this does not render the Paasche's index altogether useless. In fact, it supplements the practical utility of the Laspeyre's index. The fact that the Laspeyre's index has an upward bias and the Paasche's index downward bias, the two provide the range between which the index can vary between the base period and the given period. Interestingly, thus, the former represents the upper limit, and the latter the lower limit.

6.6.3.3 Improvements over the Laspeyre’s and Paasche’s Indices

To overcome the difficulty of overstatement of changes in prices by the Laspeyre's index and understatement by the Paasche's index, different indices have been developed to compromise and improve upon them. These are particularly useful when the given period and the base period fall quite apart and result in a greater divergence between Laspeyre's and Paasche's indices.

Other important Weighted Aggregative Indices are:

1. **Marshall-Edgeworth Index**

   The Marshall-Edgeworth Index uses the average of the base period and given period quantities/prices as the weights, and is expressed as

$$P_{0i}^{ME} = \frac{\sum p_i \left(\frac{q_0 + q_i}{2} \right)}{\sum p_0 \left(\frac{q_0 + q_i}{2} \right)} x100 \qquad \text{………… } \mathit{(6\text{-}17)}$$

$$Q_{0i}^{ME} = \frac{\sum q_i \left(\frac{p_0 + p_i}{2} \right)}{\sum q_0 \left(\frac{p_0 + p_i}{2} \right)} x100 \qquad \text{………… } \mathit{(6\text{-}18)}$$

2. **Dorbish and Bowley Index**

   The Dorbish and Bowley Index is defined as the arithmetic mean of the Laspeyre’s and Paasche’s indices.

$$P_{0i}^{DB} = \frac{P_{0i}^{La} + P_{0i}^{Pa}}{2} \qquad \text{………… } \mathit{(6\text{-}19)}$$

$$Q_{0i}^{DB} = \frac{Q_{0i}^{La} + Q_{0i}^{Pa}}{2} \qquad \text{………… } \mathit{(6\text{-}20)}$$

3. **Fisher’s Ideal Index**

The Fisher's Ideal Index is defined as the geometric mean of the Laspeyre's and Paasche's indices.

$$P_{0i}^{F} = \sqrt{P_{0i}^{La}.P_{0i}^{Pa}} \qquad \ldots\ldots\ldots\ldots(6\text{-}21)$$

$$Q_{0i}^{F} = \sqrt{Q_{0i}^{La}.Q_{0i}^{Pa}} \qquad \ldots\ldots\ldots\ldots(6\text{-}22)$$

6.6.4 INDEX OF WEIGHTED AVERAGE OF PRICE/QUANTITY RELATIVES

An alternative system of assigning weights lies in using value weights. The value weight v for any single commodity is the product of its price and quantity, that is, $v = pq$.

If the index of weighted average of price relatives is defined as

$$P_{0i} = \frac{\sum\left[v\left(\frac{p_i}{p_0}x100\right)\right]}{\sum v} \qquad \ldots\ldots\ldots\ldots(6\text{-}23)$$

then v can be obtained either as

(i) the product of the base period prices and the base period quantities denoted as v_0 that is, $v_0 = p_0 q_0$, or

(ii) the product of the base period prices and the given period quantities denoted as v_i that is, $v_i = p_0 q_i$

When v is $v_0 = p_0 q_0$, the index of weighted average of price relatives, is expressed as

$$_0P_{0i} = \frac{\sum\left[p_0q_0\left(\frac{p_i}{p_0}x100\right)\right]}{\sum p_0q_0} \qquad \ldots\ldots\ldots\ldots(6\text{-}24)$$

It may be seen that $_0P_{0i}$ is the same as the Laspeyre's aggregative price index.

Similarly, When v is $v_i = p_0 q_i$, the index of weighted average of price relatives, is expressed as

$$_iP_{0i} = \frac{\sum\left[p_0q_i\left(\frac{p_i}{p_0}x100\right)\right]}{\sum p_0q_i} \qquad \ldots\ldots\ldots\ldots(6\text{-}25)$$

It may be seen that $_iP_{0i}$ is the same as the Paasche's aggregative price index.

If the index of weighted average of quantity relatives is defined as

$$Q_{0i} = \frac{\sum\left[v\left(\frac{q_i}{q_0}x100\right)\right]}{\sum v} \qquad \ldots\ldots\ldots\ldots(6\text{-}26)$$

then v can be obtained either as

(i) the product of the base period quantities and the base period prices denoted as v_0 that is, $v_0 = q_0p_0$, or

(ii) the product of the base period quantities and the given period prices denoted as v_i that is, $v_i = q_0p_i$

When v is $v_0 = q_0p_0$, the index of weighted average of quantity relatives, is expressed as

$$_0Q_{0i} = \frac{\sum\left[q_0p_0\left(\frac{q_i}{q_0}x100\right)\right]}{\sum q_0p_0} \qquad \ldots\ldots\ldots\ldots(6\text{-}27)$$

It may be seen that $_0Q_{0i}$ is the same as the Laspeyre's aggregative quantity index.

Similarly, When v is $v_i = v_i = q_0\,p_i$, the index of weighted average of quantity relatives, is expressed as

$$_iQ_{0i} = \frac{\sum\left[q_0p_i\left(\frac{q_i}{q_0}x100\right)\right]}{\sum q_0p_i} \qquad \ldots\ldots\ldots\ldots(6\text{-}28)$$

It may be seen that $_iQ_{0i}$ is the same as the Paasche's aggregative quantity index.

Example 6-5

From the data in Example 6.2 find the:

(a) Index of Weighted Average of Price Relatives, using

(i) $v_0 = p_0 q_0$ as the value weights

(ii) $v_i = p_0 q_i$ as the value weights

(b) Index of Weighted Average of Quantity Relatives, using

(i) $v_0 = q_0 p_0$ as the value weights

(ii) $v_i = q_0 p_i$ as the value weights

Solution:

Calculations for
Index of Weighted Average of Price Relatives
(Base Year = 1980)

Item	$v_0 = p_0 q_0$	$v_1 = p_0 q_1$	$p_0 q_0\left(\frac{p_i}{p_0} x100\right)$	$p_0 q_1\left(\frac{p_1}{p_0} x100\right)$
Fish	7500	9000	1000000	1200000
Mutton	10620	11520	1357000	1472000
Chicken	9900	11000	1080000	1200000
Sum →	28020	31520	3437000	3872000

(a) Index of Weighted Average of Price Relatives, using

(i) $v_0 = p_0 q_0$ as the value weights

$$ {}_0P_{0i} = \frac{\sum\left[p_0 q_0\left(\frac{p_i}{p_0} x100\right)\right]}{\sum p_0 q_0} $$

$$ = \frac{3437000}{28020} $$

$$ = 122.66 $$

(ii) $v_i = p_0 q_i$ as the value weights

$$ {}_iP_{0i} = \frac{\sum\left[p_0 q_i\left(\frac{p_i}{p_0} x100\right)\right]}{\sum p_0 q_i} $$

$$= \frac{3872000}{31520}$$

$$= 122.84$$

Calculations for
Index of Weighted Average of Quantity Relatives
(Base Year = 1980)

Item	$v_0 = q_0 p_0$	$v_1 = q_0 p_1$	$q_0 p_0 \left(\frac{q_1}{q_0} x100 \right)$	$q_0 p_1 \left(\frac{q_1}{q_0} x100 \right)$
Fish	7500	10000	900000	1200000
Mutton	10620	13570	1152000	1472000
Chicken	9900	10800	1100000	1200000
Sum →	28020	34370	3152000	3872000

(b) Index of Weighted Average of Quantity Relatives, using

(i) $v_0 = q_0 p_0$ as the value weights

$${}_0Q_{0i} = \frac{\sum \left[q_0 p_0 \left(\frac{q_i}{q_0} x100 \right) \right]}{\sum q_0 p_0}$$

$$= \frac{3152000}{28020}$$

$$= 112.49$$

(ii) $v_i = q_0 p_i$ as the value weights

$${}_iQ_{0i} = \frac{\sum \left[q_0 p_i \left(\frac{q_i}{q_0} x100 \right) \right]}{\sum q_0 p_i}$$

$$= \frac{3872000}{34370}$$

$$= 112.66$$

Although the indices of weighted average of price/quantity relatives yield the same results as the Laspeyre's or Paasche's price/quantity indices, we do construct these indices also in situations when it is necessary and advantageous to do so. Some such situations are as follows:

(i) When a group of commodities is to be represented by a single commodity in the group, the price relative of the latter is weighted by the group as a whole.

(ii) Where the price/quantity relatives of individual commodities have been computed, these can be more conveniently utilised in constructing the index.

(iii) Price/quantity relatives serve a useful purpose in splicing two index series having different base periods.

(iv) Depersonalizing a time series requires construction of a seasonal index, which also requires the use of relatives.

6.7 TESTS OF ADEQUACY OF INDEX NUMBERS

We have discussed various formulae for the construction of index numbers. None of the formulae measures the price changes or quantity changes with perfection and has some bias. The problem is to choose the most appropriate formula in a given situation. As a measure of the formula error a number of mathematical tests, known as the *tests of consistency* or *tests of adequacy* of index number formulae have been suggested. In this section we will discuss these tests, which are also sometimes termed as the criteria for a good index number.

1. **Unit Test:** This test requires that the index number formula should be independent of the units in which the prices or quantities of various commodities are quoted. All the formulae discussed in the lesson except the index number based on Simple Aggregate of Prices/Quantities satisfy this test.
2. **Time Reversal Test:** The time reversal test, proposed by Prof Irving Fisher requires the index number formula to possess time consistency by working both forward and backward *w.r.t.* time. In his (Fisher's) words:

> *"The formula for calculating an 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 the base or putting it another way, the index number reckoned forward should be reciprocal of the one reckoned backward."*

In other words, if the index numbers are computed for the same data relating to two periods by the same formula but with the bases reversed, then the two index numbers so obtained should be the reciprocals of each other. Mathematically, we should have (omitting the factor 100),

$$P_{ab} x P_{ba} = 1 \qquad \text{…………}(6\text{-}29)$$

or more generally

$$P_{01} x P_{10} = 1 \qquad \text{…………}(6\text{-}29a)$$

Time reversal test is satisfied by the following index number formulae:

(i) Marshall-Edgeworth formula

(ii) Fisher's Ideal formula

(iii) Kelly's fixed weight formula

(iv) Simple Aggregate index

(v) Simple Geometric Mean of Price Relatives formula

(vi) Weighted Geometric Mean of Price Relatives formula with fixed weights

Lespeyre's and Pasche's index numbers do not satisfy the time reversal test.

3. **Factor Reversal Test:** This is the second of the two important tests of consistency proposed by Prof Irving Fisher. According to him:

> *"Just as our formula should permit the interchange of two times without giving inconsistent results, so it ought to permit interchanging the price and quantities without giving inconsistent results – i.e., the two results multiplied together should give the true value ratio, except for a constant of proportionality."*

This implies that if the price and quantity indices are obtained for the same data, same base and current periods and using the same formula, then their product (without the factor 100) should give the true value ratio. Symbolically, we should have (without factor 100).

$$P_{01} x Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = V_{01} \qquad \ldots\ldots\ldots\ldots(6\text{-}30)$$

Fisher's formula satisfies the factor reversal test. In fact fisher's index is the only index satisfying this test as none of the formulae discussed in the lesson satisfies this test.

Remark: Since Fisher's index is the only index that satisfies both the time reversal and factor reversal tests, it is termed as Fisher's Ideal Index.

4. **Circular Test:** Circular test, first suggested by Westergaard, is an extension of time reversal test for more than two periods and is based on the shift ability of the base period. This requires the index to work in a circular manner and this property enables us to find the index numbers from period to period without referring back to the original base each time. For three periods *a,b,c,* the test requires :

$$P_{ab} x P_{bc} x P_{ca} = 1 \qquad a \neq b \neq c \qquad \ldots\ldots\ldots\ldots(6\text{-}31)$$

In the usual notations *Eq. (6-31)* can be stated as:

$$P_{01} x P_{12} x P_{20} = 1 \qquad \ldots\ldots\ldots\ldots(6\text{-}31a)$$

For Instance

$$P_{01}^{La}\ x P_{12}^{La} x P_{21}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} x \frac{\sum p_2 q_1}{\sum p_1 q_1} x \frac{\sum p_0 q_2}{\sum p_2 q_2} \neq 1$$

Hence Laspeyre's index does not satisfy the circular test. In fact, circular test is not satisfied by any of the weighted aggregative formulae with changing weights. This test is satisfied only by the index number formulae based on:

(i) Simple geometric mean of the price relatives, and

(ii) Kelly's fixed base method

6.8 SPECIAL ISSUSES IN THE CONSTRUCTION OF INDEX NUMBERS

6.8.1 BASE SHIFTING

The need for shifting the base may arise either

(i) when the base period of a given index number series is to be made more recent, or

(ii) when two index number series with different base periods are to be compared, or

(iii) when there is need for splicing two overlapping index number series.

Whatever be the reason, the technique of shifting the base is simple:

$$\text{New Base Index Number} = \frac{\text{Old Index Number of Current Year}}{\text{Old Index Number of New Base Year}} x100$$

Example 6-6

Reconstruct the following indices using 1997 as base:

Year :	1991	1992	1993	1994	1995	1996	1997	1998
Index :	100	110	130	150	175	180	200	220

Solution:

Shifting the Base Period

Year	Index Number (1991 = 100)	Index Number (1997 = 100)
1991	100	(100/200) x100 = 50.00
1992	110	(110/200) x100 = 55.00
1993	130	(130/200) x100 = 65.00
1994	150	(150/200) x100 = 75.00
1995	175	(175/200) x100 = 87.50
1996	180	(180/200) x100 = 90.00
1997	200	(200/200) x100 = 100.00
1998	220	(220/200) x100 = 110.00

6.8.2 SPLICING TWO OVERLAPPING INDEX NUMBER SERIES

Splicing two index number series means reducing two overlapping index series with different base periods into a single series either at the base period of the old series (one with an old base year), or at the base period of the new series (one with a recent base year). This actually amounts to changing the weights of one series into the weights of the other series.

1. **Splicing the New Series to Make it Continuous with the Old Series**

   Here we reduce the new series into the old series after the base year of the former. As shown in Table 6.8.2*(i),* splicing here takes place at the base year (1980) of the new series. To do this, a ratio of the index for 1980 in the old series (200) to the index of 1980 in the new series (100) is computed and the index for each of the following years in the new series is multiplied by this ratio.

Table 6.8.2*(i)*
Splicing the New Series with the Old Series

Year	Price Index (1976 = 100) (Old Series)	Price Index (1980 = 100) (New Series)	Spliced Index Number [New Series *x* (200/100)]
1976	100	--	100
1977	120	--	120
1978	146	--	146
1979	172	--	172
1980	200	100	200
1981	--	110	220
1982	--	116	232
1983	--	125	250
1984	--	140	280

2. **Splicing the Old Series to Make it Continuous with the New Series**

This means reducing the old series into the new series before the base period of the letter. As shown in Table 6.8.2*(ii),* splicing here takes place at the base period of the new series. To do this, a ratio of the index of 1980 of the new series (100) to the index of 1980 of the old series (200) is computed and the index for each of the preceding years of the old series are then multiplied by this ratio.

Table 6.8.2*(ii)*
Splicing the Old Series with the New Series

Year	Price Index (1976 = 100) (Old Series)	Price Index (1980 = 100) (New Series)	Spliced Index Number [Old Series *x* (100/200)]
1976	100	--	50
1977	120	--	60
1978	146	--	73.50
1979	172	--	86
1980	200	100	100
1981	--	110	110
1982	--	116	116
1983	--	125	125
1984	--	140	140

6.8.3 CHAIN BASE INDEX NUMBERS

The various indices discussed so far are fixed base indices in the sense that either the base year quantities/prices (or the given year quantities/prices) are used as weights. In a dynamic situation where tastes, preferences, and habits are constantly changing, the weights should be revised on a continuous basis so that new commodities are included and the old ones deleted from consideration.

This is all the more necessary in a developing society where new substitutes keep replacing the old ones, and completely new commodities are entering the market. To take care of such changes, the base year should be the most recent, that is, the year immediately preceding the given year. This means that as we move forward, the base year should move along the given year in a chain year after year.

Conversion of Fixed-base Index into Chain-base Index

As shown in Table 6.8.3*(i),* to convert fixed-base index numbers into chain-base index numbers, the following procedure is adopted:

- The first year's index number is taken equal to 100
- For subsequent years, the index number is obtained by following formula:

$$\text{Current Year's CBI} = \frac{\text{Current Year's FBI}}{\text{Previous Year's CBI}} x\, 100$$

Table 6.8.3*(i)*
Conversion of Fixed-base Index into Chain-base Index

Year	Fixed Base Index Number (FBI)	Conversion	Chain Base Index Number (CBI)
1975	376	--	100
1976	392	(392/376) x100	104.3
1977	408	(408/392) x100	104.1
1978	380	(380/408) x100	93.1
1979	392	(392/380) x100	103.2
1980	400	(400/392) x100	102

Conversion of Chain-base Index into Fixed-base Index

As shown in Table 6.8.3*(ii),* to convert fixed-base index numbers into chain-base index numbers, the following procedure is adopted:

- The first year's index is taken what the chain base index is; but if it is to form the base it is taken equal to 100
- In subsequent years, the index number is obtained by following formula:

$$\text{Current Year's FBI} = \frac{\text{Current Year's CBI } x \text{ Previous Year's FBI}}{100}$$

Table 6.8.3*(ii)*
Conversion of Chain-base Index into Fixed-base Index

Year	Chain Base Index Number (CBI)	Conversion	Fixed Base Index Number (FBI)
1978	90	--	90
1979	120	(120 *x* 90) /100	108
1980	125	(125 *x* 108) /100	135
1981	110	(110 *x* 135) /100	148.5
1982	112	(112 *x* 148.5) /100	166.3
1983	150	(150 *x* 166.3) /100	249.45

6.9 PROBLEMS OF CONSTRUCTING INDEX NUMBERS

The above discussion enables us to identify some of the important problems, which may be faced in the construction of index numbers:

1. ***Choice of the Base Period:*** Choice of the base period is a critical decision because of its importance in the construction of index numbers. A base period is the reference period for describing and comparing the changes in prices or quantities in a given period. The selection of a base year or period does not pose difficult theoretical questions. To a large extent, the choice of the base year depends on the objective of the index. A major consideration should be to ensure that the base year is not an abnormal year. *For example,* a base period with very low price/quantity will unduly inflate, while the one with a very high figure will unduly depress, the entire index number series. An index number series constructed with any such period as the base may give very misleading results. It is, therefore, necessary that the base period be selected carefully.

   Another important consideration is that the base year should not be too remote in the past. A more recent year needs to be selected as the base year. The use of a particular year for a prolonged period would distort the changes that it purports to measure. That is why we find that the base year of major index numbers, such as consumer price index or index of industrial production, is shifted from time to time.

2. ***Selection of Weights to be Used:*** It should be amply clear from the various indices discussed in the lesson that the choice of the system of weights, which may be used, is fairly large. Since any system of weights has its own merits and is capable of giving results amenable to precise interpretations, the weights used should be decided keeping in view the purpose for which an index is constructed.

   It is also worthwhile to bear in, mind that the use of any system of weights should represent the relative importance of individual commodities that enter into the construction of an index. The interpretations that are intended to be made from an index number are also important in deciding the weights. The use of a system of weights that involves heavy computational work deserves to be avoided.

3. ***Type of Average to be Used:*** What type of average should be used is a problem specific to simple average indices. Theoretically, one can use any of the several averages that we have, such as mean, median, mode, harmonic mean, and geometric mean. Besides being locational averages, median and mode are not the appropriate averages to use especially where the number of years for which an index is to be computed, is not large.

   While the use of harmonic mean and geometric mean has some definite merits over mean, particularly when the data to be averaged refer to ratios, mean is generally more frequently used for convenience in computations.

4. ***Choice of Index:*** The problem of selection of an appropriate index arises because of availability of different types of indices giving different results when applied to the same data. Out of the various indices discussed, the choice should be in favour of one

which is capable of giving more accurate and precise results, and which provides answer to specific questions for which an index is constructed.

While the Fisher's index may be considered ideal for its ability to satisfy the tests of adequacy, this too suffers from two important drawbacks. First, it involves too lengthy computations, and second, it is not amenable to easy interpretations as are the Laspeyre's and Paasche's indices. The use of the term ideal does not, however, mean that it is the best to use under all types of situations. Other indices are more appropriate under situations where specific answers are needed.

5. ***Selection of Commodities:*** Commodities to be included in the construction of an index should be carefully selected. Only those commodities deserve to be included in the construction of an index as would make it more representative. This, in fact, is a problem of sampling, for being related to the selection of commodities to be included in the sample.

   In this context, it is important to note that the selection of commodities must not be based on random sampling. The reason being that in random sampling every commodity, including those that are not important and relevant, have equal chance of being selected, and consequently, the index may not be representative. The choice of commodities has, therefore, to be deliberate and in keeping with the relevance and importance of each individual commodity to the purpose for which the index is constructed.

6. ***Data Collection:*** Collection of data through a sample is the most important issue in the construction of index numbers. The data collected are the raw material of an index. Data quality is the basic factor that determines the usefulness of an index. The data have to be as accurate, reliable, comparable, representative, and adequate, as possible.

The practical utility of an index also depends on how readily it can be constructed. Therefore, data should be collected from where these can be easily available. While the purpose of an index number will indicate what type of data are to be collected, it also determines the source from where the data can be available.