MAR GREGORIOS COLLEGE OF ARTS & SCIENCE

Block No.8, College Road, Mogappair West, Chennai – 37

Affiliated to the University of Madras
Approved by the Government of Tamil Nadu
An ISO 9001:2015 Certified Institution

DEPARTMENT OF COMMERCE (CORPORATE SECRETARYSHIP)

SUBJECT NAME: BUSINESS STATISTICS

SUBJECT CODE: AY33A

SEMESTER: III

PREPARED BY: PROF.RADHIKA

UNIT – I

Introduction

Meaning and Definition of Statistics:

Meaning:

"Statistics", that a word is often used, has been derived from the Latin word 'Status' that means a group of numbers or figures; those represent some information of our human interest.

We find statistics in everyday life, such as in books or other information papers or TV or newspapers.

Although, in the beginning it was used by Kings only for collecting information about states and other information which was needed about their people, their number, revenue of the state etc.

This was known as the science of the state because it was used only by the Kings. So it got its development as 'Kings' subject or 'Science of Kings' or we may call it as "Political Arithmetic's". It was for the first time, perhaps in Egypt to conduct census of population in 3050 B.C. because the king needed money to erect pyramids. But in India, it is thought, that, it started dating back to Chandra Gupta Maurya's kingdom under Chankya to collect the data of births and deaths. TM has also been stated in Chankya'sArthshastra.

But now-a-days due to its pervading nature, its scope has increased and widened. It is now used in almost in all the fields of human knowledge and skills like Business, Commerce, Economics, Social Sciences, Politics, Planning, Medicine and other sciences, Physical as well as Natural.

Definition:

The term 'Statistics' has been defined in two senses, i.e. in Singular and in Plural sense.

"Statistics has two meanings, as in plural sense and in singular sense".

—Oxford Dictionary

In plural sense, it means a systematic collection of numerical facts and in singular sense; it is the science of collecting, classifying and using statistics.

A. In the Plural Sense:

"Statistics are numerical statements of facts in any department of enquiry placed in relation to each other." —A.L. Bowley

"The classified facts respecting the condition of the people in a state—especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement." —Webster

These definitions given above give a narrow meaning to the statistics as they do not indicate its various aspects as are witnessed in its practical applications. From the this point of view the definition given by Prof. Horace Sacrist appears to be the most comprehensive and meaningful:

"By statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose, and placed in relation to each other."—Horace Sacrist

B. In the Singular Sense:

"Statistics refers to the body of technique or methodology, which has been developed for the collection, presentation and analysis of quantitative data and for the use of such data in decision making." —Ncttor and Washerman

"Statistics may rightly be called the science of averages." —Bowleg

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

Stages of Investigations:

1. Collection of Data:

It is the first stage of investigation and is regarding collection of data. It is determined that which method of collection is needed in this problem and then data are collected.

2. Organisation of Data:

It is second stage. The data are simplified and made comparative and are classified according to time and place.

3. Presentation of Data:

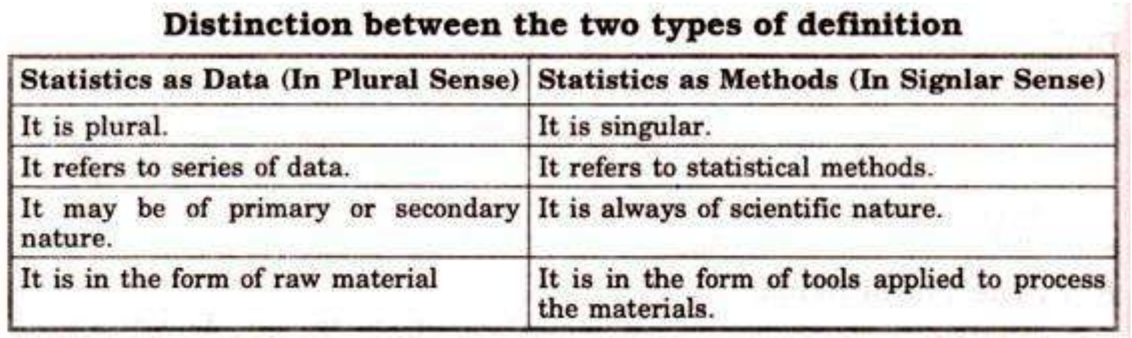
In this third stage, organised data are made simple and attractive. These are presented in the form of tables diagrams and graphs.

4. Analysis of Data:

Forth stage of investigation is analysis. To get correct results, analysis is necessary. It is often undertaken using Measures of central tendencies, Measures of dispersion, correlation, regression and interpolation etc.

5. Interpretation of Data:

In this last stage, conclusions are enacted. Use of comparisons is made. On this basis, forecasting is made.



Distinction between the two types of definition

Statistics as Data (In Plural Sense)	Statistics as Methods (In Signlar Sense)
It is plural.	It is singular.
It refers to series of data.	It refers to statistical methods.
It may be of primary or secondary nature.	It is always of scientific nature.
It is in the form of raw material	It is in the form of tools applied to process the materials.

Some Modern Definitions:

From the above two senses of statistics, modem definitions have emerged as given below:

"Statistics is a body of methods for making wise decisions on the face of uncertainty." —Wallis and Roberts

"Statistics is a body of methods for obtaining and analyzing numerical data in order to make better decisions in an uncertain world." —Edward N. Dubois

So, from above definitions we find that science of statistics also includes the methods of collecting, organising, presenting, analysing and interpreting numerical facts and decisions are taken on their basis.

The most proper definition of statistics can be given as following after analysing the various definitions of statistics.

"Statistics in the plural sense are numerical statements of facts capable of some meaningful analysis and interpretation, and in singular sense, it relates to the collection, classification, presentation and interpretation of numerical data."

COLLECTION OF DATA, CLASSIFICATION AND TABULATION

Introduction:

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

The following are some of the **objectives of collecting statistical information**.

1. To describe the methods of collecting primary statistical information.

2. To consider the status involved in carrying out a survey.

3. To analyse the process involved in observation and interpreting.

4. To define and describe sampling.

5. To analyse the basis of sampling.

6. To describe a variety of sampling methods.

Statistical investigation is a comprehensive and requires systematic collection of data about some group of people or objects, describing and organizing the data, analyzing the data with 28 the help of different statistical method, summarizing the analysis and using these results for making judgements, decisions and predictions. The validity and accuracy of final judgement is most crucial and depends heavily on how well the data was collected in the first place. The quality of data will greatly affect the conditions and hence at most importance must be given to this process and every possible precautions should be taken to ensure accuracy while collecting the data.

Nature of data:

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

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

Time series data:

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

Spatial Data:

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

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

Categories of data:

Any statistical data can be classified under two categories depending upon the sources utilized. These categories are, 1. Primary data and2. Secondary data

Primary data:

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

Example 4: If a researcher is interested to know the impact of noonmeal scheme for the school children, he has to undertake a survey and collect data on the opinion of parents and children by asking relevant questions. Such a data collected for the purpose is called primary data.

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

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

1. Direct personalinterviews:

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

2. Indirect OralInterviews:

Under this method the investigator contacts witnesses or neighbours or friends or some other third parties who are capable of supplying the necessary information. This method is preferred if the required information is on addiction or cause of fire or theft or murder etc., If a fire has broken out a certain place, the persons living in neighbourhood and witnesses are likely to give information on the cause of fire. In some cases, police interrogated third parties who are supposed to have knowledge of a theft or a murder and get some clues. Enquiry committees appointed by governments generally adopt this method and get people' s views and all possible details of facts relating to the enquiry. This method is suitable whenever direct sources do not exists or cannot be relied upon or would be unwilling to part with the information. The validity of the results depends upon a few factors, such as the nature of the person whose evidence is being recorded, the ability of the interviewer to draw out information from the third 32 parties by means of appropriate questions and cross examinations, and the number of persons interviewed. For the success of this method one person or one group alone should not be relied upon.

3. Information fromcorrespondents:

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

4. Mailed questionnaire method:

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

5. Schedules sent through Enumerators:

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

Secondary Data:

Secondary data are those data which have been already collected and analysed by some earlier agency for its own use; and later the same data are used by a different agency.

According to W.A.Neiswanger, ‘ A primary source is a publication in which the data are published by the same authority which gathered and analysed them. A secondary source is a publication, reporting the data which have been gathered by other authorities and for which others are responsible’.

Sources of Secondary data:

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

The sources of secondary data can broadly be classified under two heads:

1. Published sources, and 2. Unpublished sources.

1. Published Sources: The various sources of published data are: Clinical and other personal records, death certificates, published mortality statistics, census publications, etc. Examples include:

1. Official publications of Central Statistical Authority

2. Publication of Ministry of Health and Other Ministries

3. News Papers and Journals.

4. International Publications like Publications by WHO, World Bank, UNICEF

5. Records of hospitals or any Health Institutions.

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

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

1. How the data has been collected and processed.

2. The accuracy of the data.

3. How far the data has been summarized.

4. How comparable the data is with other tabulations.

5. How to interpret the data, especially when figures collected for one purpose is used for another Generally speaking, with secondary data, people have to compromise between what they want and what they are able to find.

Classification:

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

For Example:

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

Objects of Classification: The following are main objectives of classifying the data:

1. It condenses the mass of data in an easily assumable form.

2. It eliminates unnecessary details.

3. It facilitates comparison and highlights the significant aspect of data.

4. It enables one to get a mental picture of the information and helps in drawing inferences.

5. It helps in the statistical treatment of the information collected.

Types of classification:

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

a) Chronological classification

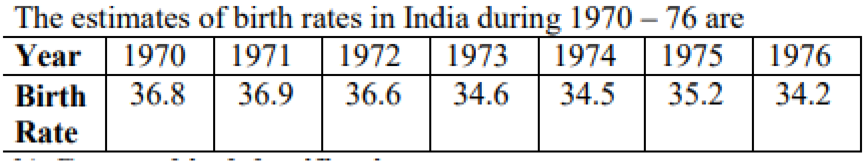
b) Geographical classification

c) Qualitative classification

d) Quantitative classification

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

Eg:

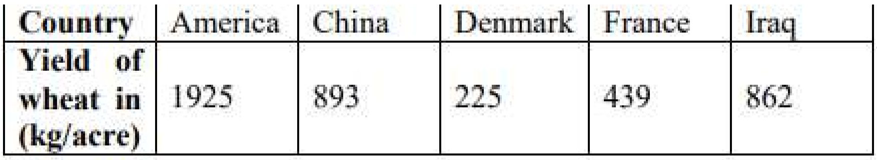


The estimates of birth rates in India during 1970 – 76 are

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

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

Eg:



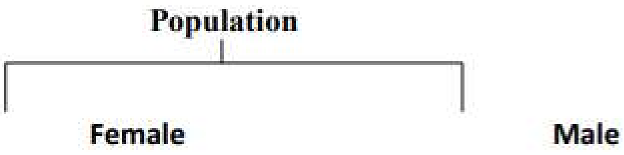
Country	America	China	Denmark	France	Iraq
Yield of wheat in (kg/acre)	1925	893	225	439	862

c) Qualitative classification:

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

For example, if the population to be classified in respect to one attribute, say sex, then we can classify them into two namely that of males and females.

Similarly, they can also be classified into 'married or ' single' on the basis of another attribute 'marital status'. Thus when the classification is done with respect to one attribute, which is dichotomous in nature, two classes are formed, one possessing the attribute and the other not possessing the attribute. This type of classification is called simple or dichotomous classification. A simple classification may be shown as under



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

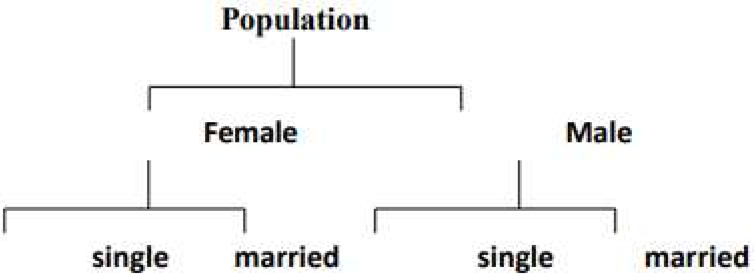
(i) Male married

(ii) Male single

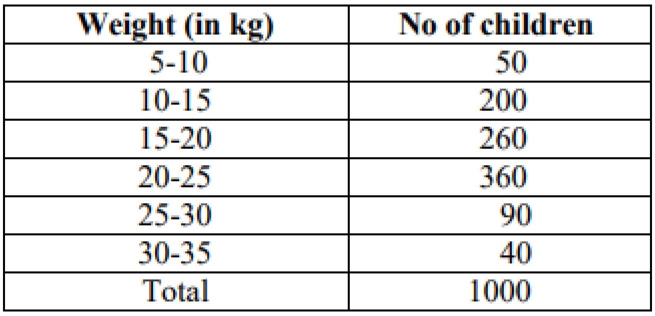
(iii) Female married

(iv) Female single.

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



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



Weight (in kg)	No of children
5-10	50
10-15	200
15-20	260
20-25	360
25-30	90
30-35	40
Total	1000

In this type of classification there are two elements, namely

(i) the variable (i.e) the weight in the above example,

and (ii) the frequency in the number of children.

There are 50 childre having weights ranging from 5 to 10 kg, 200 children. having weight ranging between 10 to 15 kg and so on. 3.5

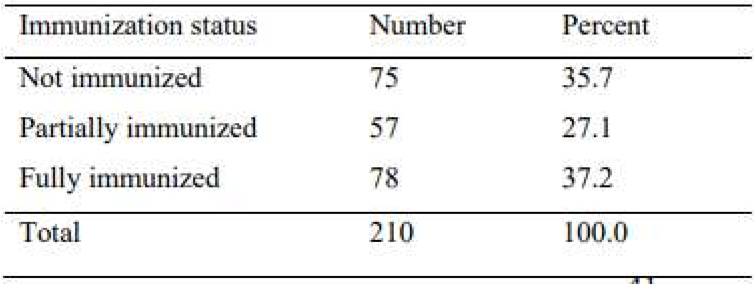
Tabulation:

Tabulation is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information. A table is a systematic arrangement of classified data in columns and rows.

Thus, a statistical table makes it possible for the investigator to present a huge mass of data in a detailed and orderly form. It facilitates comparison and often reveals certain patterns in data which are otherwise not obvious Classification and ' Tabulation' , as a matter of fact, are not two distinct processes. Actually they go together.

Before tabulation data are classified and then displayed under different columns and rows of table.

Table : Overall immunization status of children in AdamiTullu Woreda, Feb. 1995



Immunization status	Number	Percent
Not immunized	75	35.7
Partially immunized	57	27.1
Fully immunized	78	37.2
Total	210	100.0

Data Presentation:

Data can be presented in one of the three ways:
– as text;
– in tabular form; or
– in graphical form.

Methods of presentation must be determined according to the data format, the method of analysis to be used, and theinformation to be emphasized. Inappropriately presented data fail to clearly convey information to readers andreviewers.Even when the same information is being conveyed, different methods of presentation must be employed depending on what specific information is going to be emphasized. A method of presentation must be chosen after carefully weighing the advantages and disadvantages of different methods of presentation. For easy comparison of different methods of presentation, let us look at a table (Table 1) and a line graph (Fig. 1) that present the same information [1]. If one wishes to compare or introduce two values at a certain time point, it is appropriate to use text or the written language. However, a table is the most appropriate when all information requires equal attention, and it allows readers to selectively look at information of their own interest. Graphs allow readers to understand the overall trend in data, and intuitively understand the comparison results between two groups. One thing to always bear in mind regardless of what method is used.

PRESENTATION OF DATA:

The main portion of Statistics is the display of summarized data. Data is initially collected from a given source, whether they are experiments, surveys, or observation, and is presented in one of four methods:

Textual Method

The reader acquires information through reading the gathered data.

Tabular Method

Provides a more precise, systematic and orderly presentation of data in rows or columns.

Semi-tabular Method

Uses both textual and tabular methods.

Graphical Method

The utilization of graphs is most effective method of visually presenting statistical results or findings.

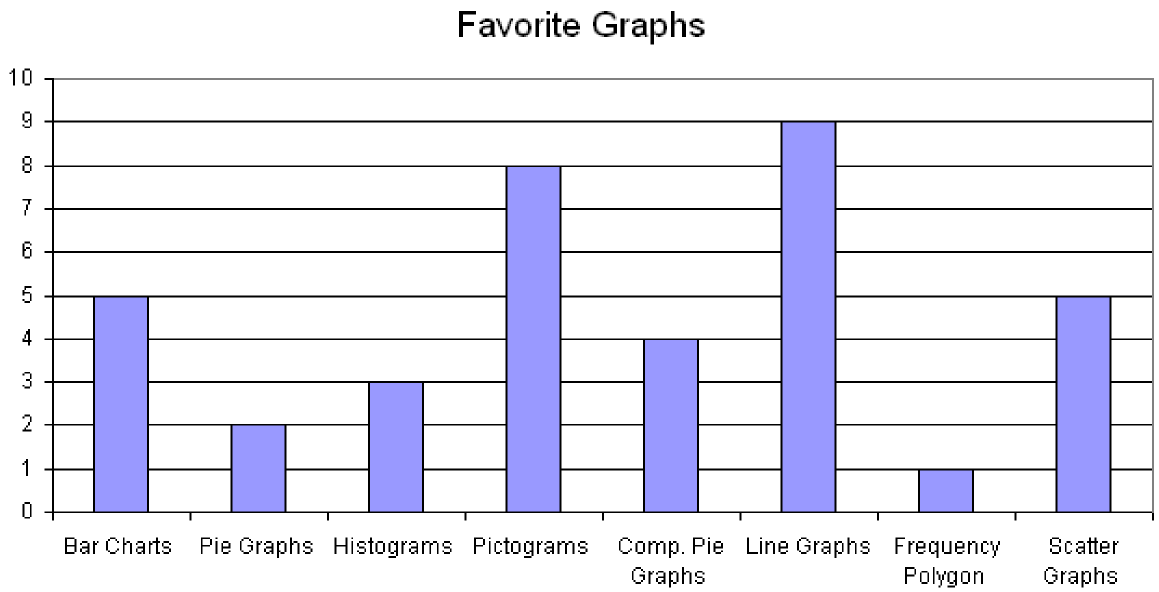
The Bar Chart:

The Bar Chart (or Bar Graph) is one of the most common ways of displaying catagorical/qualitative data. Bar Graphs consist of 2 variables, one response (sometimes called "dependent") and one predictor (sometimes called "independent"), arranged on the horizontal and vertical axis of a graph. The relationship of the predictor and response variables is shown by a mark of some sort (usually a rectangular box) from one variable's value to the other's.

To demonstrate we will use the following data representing a hypothetical relationship between a qualitative predictor variable, "Graph Type", and a quantitative response variable, "Votes".

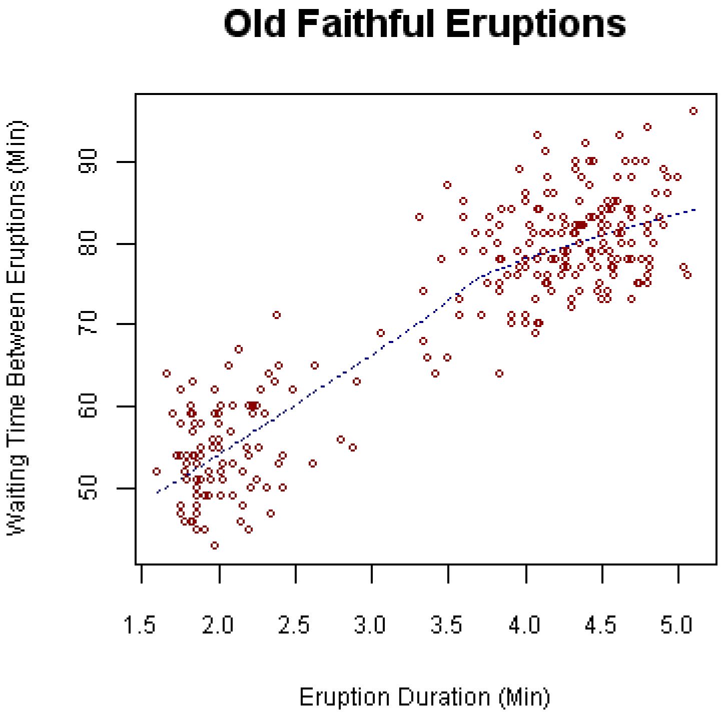
Graph Type	**Votes**
Bar Charts	5
Pie Graphs	2
Histograms	3
Pictograms	8
Comp. Pie Graphs	4
Line Graphs	9
Frequency Polygon	1
Scatter Graphs	5

From this data we can now construct an appropriate graphical representation which, in this case will be a Bar Chart. The graph may be orientated in several ways, of which the vertical chart is most common, with the horizontal chart also being used often.



Scatter Plot:

Scatter Plot is used to show the relationship between 2 numeric variables. It is not useful when comparing discrete variables versus numeric variables. A scatter plot matrix is a collection of pairwise scatter plots of numeric variables.



A Pie-Chart:

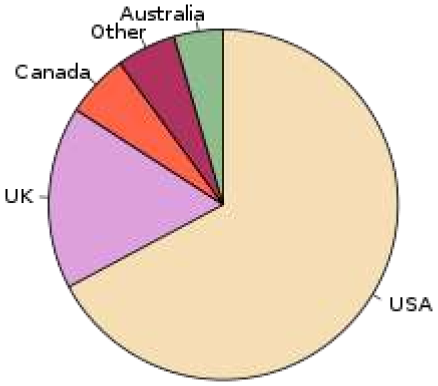
A Pie-Chart/Diagram is a graphical device - a circular shape broken into sub-divisions. The sub-divisions are called "*sectors*", whose areas are proportional to the various parts into which the whole quantity is divided. The sectors may be coloured differently to show the relationship of parts to the whole. A pie diagram is an alternative of the sub-divided bar diagram.

To construct a pie-chart, first we draw a circle of any suitable radius then the whole quantity which is to be divided is equated to 360 degrees. The different parts of the circle in terms of angles are calculated by the following formula.

```
Component Value / Whole Quantity * 360
```

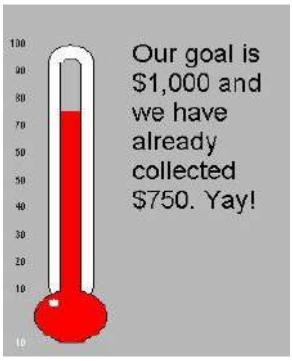
The component parts i.e. sectors have been cut beginning from top in clockwise order.

Note that the percentages in a list may not add up to exactly 100% due to rounding. For example if a person spends a third of their time on each of three activities: 33%, 33% and 33% sums to 99%.



Pictogram:

A pictogram is simply a picture that conveys some statistical information. A very common example is the thermometer graph so common in fund drives. The entire thermometer is the goal (number of dollars that the fund raisers wish to collect. The red stripe (the "mercury") represents the proportion of the goal that has already been collected.



Line graph:

Basically, a **line graph** can be, for example, a picture of what happened by/to something (a variable) during a specific time period (also a variable).

On the left side of such a graph usually is as an indication of that "something" in the form of a scale, and at the bottom is an indication of the specific time involved.

Usually a **line graph** is plotted after a table has been provided showing the relationship between the two variables in the form of pairs. Just as in (x,y) graphs, each of the pairs results in a specific point on the graph, and being a LINE graph these points are connected to one another by a LINE.

Many other line graphs exist; they all CONNECT the points by LINEs, not necessarily straight lines. Sometimes polynomials, for example, are used to describe approximately the basic relationship between the given pairs of variables, and between these points. The higher the degree of the polynomial, the more accurate is the "picture" of that relationship, but the degree of that polynomial must never be higher than *n-1*, where *n* is the number of the given points.

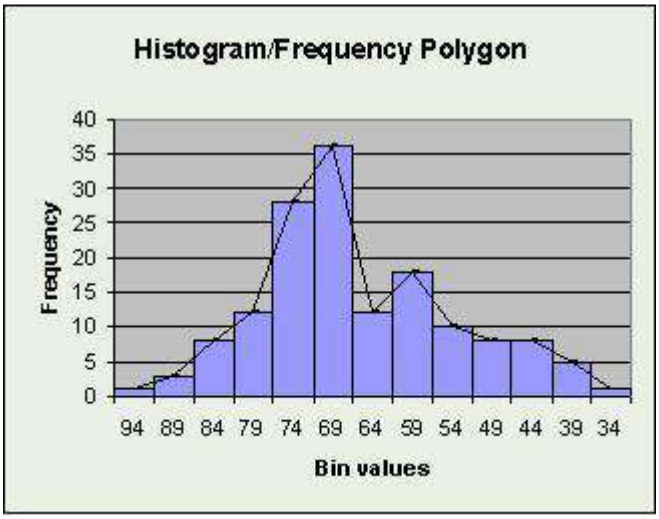
Frequency Polygan:

Midpoints of the interval of corresponding rectangle in a histogram are joined together by straight lines. It gives a polygon i.e. a figure with many angles.

It is used when two or more sets of data are to be illustrated on the same diagram such as death rates in smokers and non-smokers, birth and death rates of a population etc.

One way to form a frequency polygon is to connect the midpoints at the top of the bars of a histogram with line segments (or a smooth curve). Of course the midpoints themselves could easily be plotted without the histogram and be joined by line segments. Sometimes it is beneficial to show the histogram and frequency polygon together.But sometimes, the frequency polygon is much more accurate than the histogram because you can evaluate which is the low point and the high point.

Unlike histograms, frequency polygons can be superimposed so as to compare several frequency distributions.



UNIT – II

Measure of central tendency

A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. You can think of it as the tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.

Choosing the best measure of central tendency depends on the type of data you have.

Mean

The mean is the arithmetic average, and it is probably the measure of central tendency that you are most familiar. Calculating the mean is very simple. You just add up all of the values and divide by the number of observations in your dataset.

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

The calculation of the mean incorporates all values in the data. If you change any value, the mean changes. However, the mean doesn't always locate the center of the data accurately.

The **mean** of a set of observations is the average. It is obtained by dividing the sum of data by the number of observations.

The formula is:

$$\text{Mean} = \frac{\text{Sum of data}}{\text{Number of observations}}$$

Example:

Find the mean of the following set of integers.

8, 11, –6, 22, –3

Solution:

$$\text{Mean} = \frac{8+11+(-6)+22+(-3)}{5} = 6.4$$

When there are changes in the number or the values of the observations in a set, the mean will be changed.

Example:

The mean score of a group of 20 students is 65. Two other students whose scores are 89 and 85 were added to the group. What is the new mean of the group of students?

Solution:

The formula:

$$\text{Mean} = \frac{\text{Total score}}{\text{Number of students}}$$

can rewritten as:

Total score = Mean × Number of students

Total score of the original group = 65 × 20 = 1,300

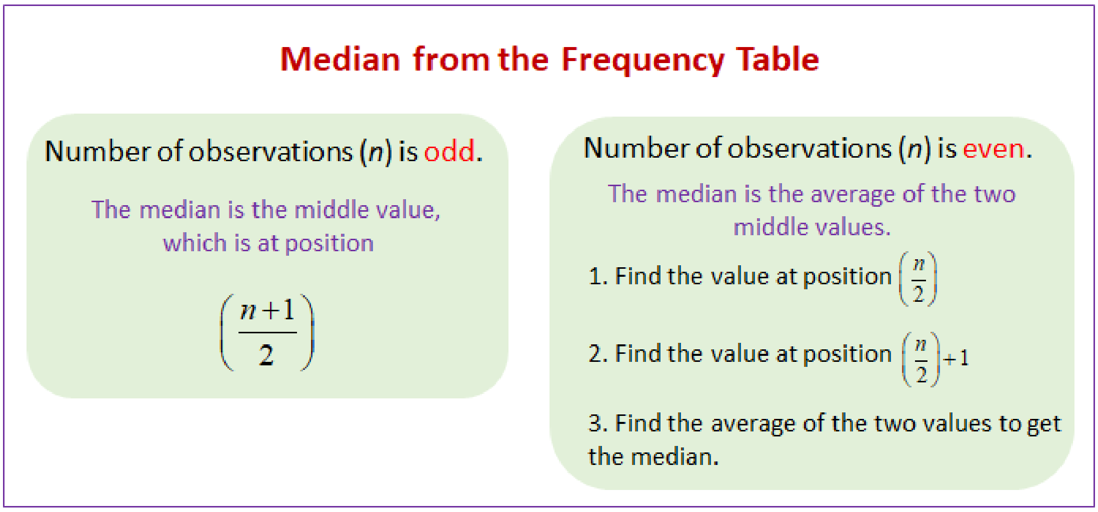
Total score of the new group
= Total score of the original group + scores of the 2 new students
= 1,300 + 89 + 85 = 1,474
Number of students in the new group
= Number of students in the original group + Number of new data
= 20 + 2 = 22

$$\text{Mean of the new group} = \frac{1{,}474}{22} = 67$$

Median:

The median is the **middle value** in an ordered set of data. In a frequency table, the observations are already arranged in an ascending order. We can obtain the median by looking for the value in the middle position. If there is an odd number of observations, the median is the middle number. If there is an even number of observations, the median will be the mean of the two central numbers.

The following table shows how to find the median from the frequency table with odd number of observations and with even number of observations. Scroll down the page for examples and step-by-step solutions.



Median from the Frequency Table

Number of observations (n) is odd.

The median is the middle value, which is at position

$$\left(\frac{n+1}{2}\right)$$

Number of observations (n) is even.

The median is the average of the two middle values.

1. Find the value at position $\left(\frac{n}{2}\right)$
2. Find the value at position $\left(\frac{n}{2}\right)+1$
3. Find the average of the two values to get the median.

How To Find The Median Of A Frequency Table When The Number Of Observations Is Odd?

Case 1. When the number of observations (n) is odd, then the median is the value at the $\left(\frac{n+1}{2}\right)^{\text{th}}$ position.

Example:

The following is a frequency table of the score obtained in a mathematics quiz. Find the median score.

Score	0	1	2	3	4
Frequency	3	4	7	6	3

Solution:

Number of scores = 3 + 4 + 7 + 6 + 3 = 23 (odd number)

Since the number of scores is odd, the median is at the $\left(\frac{n+1}{2}\right)^{th} = \left(\frac{23+1}{2}\right)^{th} = 12^{th}$ position.

To find out the 12 th position, we need to add up the frequencies as shown:

Score	0	1	2	3	4
Frequency	3	4	7	6	3
Position	3	3 + 4 = 7	7 + 7 =14		

The 12th position is after the 7th position but before the 14th position. So, the median is 2.

How To Find The Median Of A Frequency Table When The Number Of Observations Is Even?

Case 2. When *the number of observations (n) is even, then the median is the average of values at the n/2 and (n/2 + 1) positions.*

Example:

The table is a frequency table of the scores obtained in a competition. Find the median score.

Scores	0	1	2	3	4
Frequency	11	9	5	10	15

Solution:

Number of scores = 11 + 9 + 5 + 10 + 15 = 50 (even number).Since the number of scores is even, the median is at the average of the $\left(\frac{n}{2}\right)^{th} = \left(\frac{50}{2}\right)^{th} = 25^{th}$ position and $\left(\frac{n}{2}+1\right)^{th} = 26^{th}$ position.

To find out the 25th position and 26th position, we add up the frequencies as shown:

Scores	0	1	2	3	4
Frequency	11	9	5	10	15
Position	11	11 + 9 = 20	20 + 5 = 25	25 + 10 = 35	36 to 50

The score at the 25th position is 2 and the score at the 26th position is 3.

The median is the average of the scores at 25th and 26th positions = $\frac{2+3}{2} = 2.5$

Mode

The mode of a set of observations is the value that occurs most frequently in the set. A set of observations may have no mode, one mode or more than one mode.

Example:

Find the mode of the following set of scores.

14 11 15 9 11 15 11 7 13 12

Solution:

The mode is 11 because 11 occurred more times than the other numbers.

If the observations are given in the form of a frequency table, the mode is the value that has the highest frequency.

Example:

Find the mode of the following set of marks.

Marks	1	2	3	4	5
Frequency	6	7	7	5	3

Solution:

The marks 2 and 3 have the highest frequency. So, the modes are 2 and 3.

Note: The above example shows that a set of observations may have more than one mode.

Example:

Find the mode for each of the following frequency tables: The frequency table below shows the weights of different bags of rice.

Weight (kg)	45	50	55	60	65	70	75	80
Bags of rice (Frequency)	8	11	7	10	9	10	12	8

There are 8 number cards with values 0 – 7. Each time a card is drawn at random and the card value is recorded. The frequency refers to the number of times a value is shown.

Card values	0	1	2	3	4	5	6	7
Frequency	8	12	7	10	12	13	12	10

Solution:

a) Mode: 75 kg (highest frequency of 12)

b) Mode: 5 (highest frequency of 13)

Example:

The following frequency table shows the marks obtained by students in a quiz. Given that 4 is the mode, what is the least value for x?

Marks	1	2	3	4	5	6
Number of students (Frequency)	7	9	10	x	9	11

Solution:

x is as least 12

(if x is less than 12 then 4 will not be the mode)

The Harmonic Mean:

The harmonic mean is the reciprocal of the average of the reciprocals/

Reciprocal just means 1*1*value.

The formula is:

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \ldots}$$

Where a,b,c,... are the values, and n is how many values.

Steps:

- Calculate the reciprocal (1/value) for every value.
- Find the average of those reciprocals (just add them and divide by how many there are)
- Then do the reciprocal of that average (=1/average)

Example: What is the harmonic mean of 1, 2 and 4?

The reciprocals of 1, 2 and 4 are:

*1*11 = 1,   1*1*2 = 0.5,   1*1*4 = 0.25

Now add them up:

1 + 0.5 + 0.25 = 1.75

Divide by how many:

Average = *1.75*3

The reciprocal of that average is our answer:

Harmonic Mean = *3/*1.75 = 1.714 (to 3 places)

In *some* rate type questions the harmonic mean gives the true answer!

Example: we travel 10 km at 60 km/h, than another 10 km at 20 km/h, what is our average speed?

Harmonic mean = 2/(*1*60 + *1*20) = 30 km/h

Check: the 10 km at 60 km/h takes 10 minutes, the 10 km at 20 km/h takes 30 minutes, so the total 20 km takes 40 minutes, which is 30 km per hour.

The Geometric Mean:

Definition:

For **n** numbers: multiply them all together and then take the nth root (written $^n\sqrt{\ }$). More formally, the geometric mean of **n** numbers $\mathbf{a_1}$ **to** $\mathbf{a_n}$ is:

$^n\sqrt{(a_1 \times a_2 \times \ldots \times a_n)}$.

The Geometric Mean is a special type of average where we multiply the numbers together and then take a square root (for two numbers), cube root (for three numbers) etc.

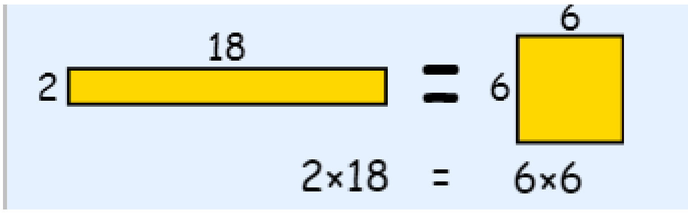
Example: What is the Geometric Mean of 2 and 18?

- First we multiply them: 2 × 18 = 36
- Then (as there are two numbers) take the square root: $\sqrt{36}$ = **6.**

In one line:

Geometric Mean of 2 and 18 = $\sqrt{(2 \times 18)} = 6$

It is like the area is the same!



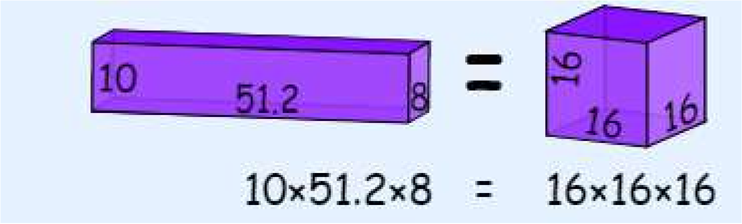
Example: What is the Geometric Mean of 10, 51.2 and 8?

- First we multiply them: 10 × 51.2 × 8 = 4096
- Then (as there are three numbers) take the cube root: $^3\sqrt{4096}$ = **16**

In one line:

Geometric Mean = $^3\sqrt{(10 \times 51.2 \times 8)} = 16$

It is like the volume is the same:



Measure of Variation:

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The standard deviation is a number that measures how far data values are from their mean.

Standard deviation

The standard deviation is the average amount of variability in your dataset.

It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

There are six steps for finding the standard deviation by hand:

1. List each score and find their mean.
2. Subtract the mean from each score to get the deviation from the mean.
3. Square each of these deviations.
4. Add up all of the squared deviations.
5. Divide the sum of the squared deviations by *n* – 1 (for a sample) or *N* (for a population).
6. Find the square root of the number you found.

Standard deviation example

Step 1: Data (minutes)	**Step 2: Deviation from mean**	**Steps 3 + 4: Squared deviation**
72	72 – 207.5 = -135.5	18360.25
110	110 – 207.5 = -97.5	9506.25
134	134 – 207.5 = -73.5	5402.25
190	190 – 207.5 = -17.5	306.25
238	238 – 207.5 = 30.5	930.25
287	287 – 207.5 = 79.5	6320.25
305	305 – 207.5 = 97.5	9506.25
324	324 – 207.5 = 116.5	13572.25
Mean = **207.5**	Sum = 0	Sum of squares = **63904**

Because you're dealing with a sample, you use *n* – 1.
n – 1 = **7**

63904 / 7 = **9129.14**

s = √9129.14 = 95.54

The standard deviation of your data is **95.54**. This means that on average, each score deviates from the mean by 95.54 points.

Standard deviation formula for populations

If you have data from the entire population, use the population standard deviation formula:

Explanation

Formula

$$\sigma = \sqrt{\frac{\Sigma\,(X - \mu)^2}{N}}$$

- σ = population standard deviation
- ∑ = sum of…
- *X* = each value
- μ = population mean
- *N* = number of values in the population

Standard deviation formula for samples

If you have data from a sample, use the sample standard deviation formula:

Formula **Explanation**

$$s = \sqrt{\frac{\Sigma\,(X - \bar{x})^2}{n - 1}}$$

- *s* = sample standard deviation
- ∑ = sum of…
- *X* = each value
- $\bar{x}$ = sample mean
- *n* = number of values in the sample

Why use *n* – 1 for sample standard deviation?

Samples are used to make statistical inferences about the population that they came from.

When you have population data, you can get an exact value for population standard deviation. Since you collect data from every population member, the standard deviation reflects the precise amount of variability in your distribution, the population.

But when you use sample data, your sample standard deviation is always used as an estimate of the population standard deviation. Using *n* in this formula tends to give you a biased estimate that consistently underestimates variability.

Reducing the sample *n* to *n* – 1 makes the standard deviation artificially large, giving you a conservative estimate of variability.

While this is not an unbiased estimate, it is a less biased estimate of standard deviation: it is better to overestimate rather than underestimate variability in samples.

The difference between biased and conservative estimates of standard deviation gets much smaller when you have a large sample size.

Mean Deviation:

In statistics and mathematics, the deviation is a measure that is used to find the difference between the observed value and the expected value of a variable. In simple words, the deviation is the distance from the center point. Similarly, the mean deviation is used to calculate how far the values fall from the middle of the data set. In this article, let us discuss the definition, formula, and examples in detail.

Mean Deviation Definition:

The mean deviation is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set. The mean deviation of the data values can be easily calculated using the below procedure.

Step 1: Find the mean value for the given data values

Step 2: Now, subtract the mean value from each of the data values given (Note: Ignore the minus symbol)

Step 3: Now, find the mean of those values obtained in step 2.

Mean Deviation Formula:

The formula to calculate the mean deviation for the given data set is given below.

Mean Deviation = $[\Sigma |X - \mu|]/N$

Here,

Σ represents the addition of values

X represents each value in the data set

μ represents the mean of the data set

N represents the number of data values

$|\ |$ represents the absolute value, which ignores the “-” symbol

Mean Deviation for Frequency Distribution:

To present the data in the more compressed form we group it and mention the frequency distribution of each such group. These groups are known as class intervals.

Grouping of data is possible in two ways:

1. Discrete Frequency Distribution
2. Continuous Frequency Distribution

In the upcoming discussion, we will be discussing mean absolute deviation in a discrete frequency distribution.

Let us first know what is actually meant by the discrete distribution of frequency.

Mean Deviation for Discrete Distribution Frequency:

As the name itself suggests, by discrete we mean distinct or non-continuous. In such a distribution the frequency (number of observations) given in the set of data is discrete in nature.

If the data set consists of values x_1,x_2, x_3……..x_n each occurring with a frequency of f_1, f_2… f_n respectively then such a representation of data is known as the discrete distribution of frequency.

To calculate the mean deviation for grouped data and particularly for discrete distribution data the following steps are followed:

Step I: The measure of central tendency about which mean deviation is to be found out is calculated. Let this measure be a.

If this measure is mean then it is calculated as,

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

$$\Rightarrow \bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i$$

where N=∑ni=1fi
If the measure is median then the given set of data is arranged in ascending order and then the cumulative frequency is calculated then the observations whose cumulative frequency is equal to or just greater than N/2 is taken as the median for the given discrete distribution of frequency and it is seen that this value lies in the middle of the frequency distribution.

Step II: Calculate the absolute deviation of each observation from the measure of central tendency calculated in step (I)

StepIII: The mean absolute deviation around the measure of central tendency is then calculated by using the formula

$$M.A.D(a) = \frac{\sum_{i=1}^{n} f_i \left|x_i - a\right|}{N}$$

If the central tendency is mean then,

$$M.A.D(\bar{x}) = \frac{\sum_{i=1}^{n} f_i\,|x_i - \bar{x}|}{N}$$

In case of median

$$M.A.D(M) = \frac{\sum_{i=1}^{n} f_i\,|x_i - M|}{N}$$

Let us look into the following examples for a better understanding.

Read: Mean Deviation for Continous Frequency Distribution

Mean Deviation Examples

Example 1:

Determine the mean deviation for the data values 5, 3,7, 8, 4, 9.

Solution:

Given data values are 5, 3, 7, 8, 4, 9.

We know that the procedure to calculate the mean deviation.

First, find the mean for the given data:

Mean, µ = (5+3+7+8+4+9)/6

µ = 36/6

µ = 6

Therefore, the mean value is 6.

Now, subtract each mean from the data value, and ignore the minus symbol if any

(Ignore”-”)

5 – 6 = 1

3 – 6 = 3

7 – 6 = 1

8 – 6 = 2

4 – 6 = 2

9 – 6 = 3

Now, the obtained data set is 1, 3, 1, 2, 2, 3.

Finally, find the mean value for the obtained data set

Therefore, the mean deviation is

= (1+3 + 1+ 2+ 2+3) /6

= 12/6

= 2

Hence, the mean deviation for 5, 3,7, 8, 4, 9 is 2.

Example 2:

In a foreign language class, there are 4 languages, and the frequencies of students learning the language and the frequency of lectures per week are given as:

Language	Sanskrit	Spanish	French	English
No. of students(x_i)	6	5	9	12
Frequency of lectures(f_i)	5	7	4	9

Calculate the mean deviation about the mean for the given data.

Solution: The following table gives us a tabular representation of data and the calculations

x_i	f_i	$x_i f_i$	$\lvert x_i - \bar{x}\rvert$	$f_i\lvert x_i - \bar{x}\rvert$
6	5	30	2.36	11.8
5	7	35	3.36	23.52
9	4	36	0.64	2.56
12	9	108	3.64	32.76
	$\sum f_i = 25$	$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i = 8.36$		$\sum_{i=1}^{n} f_i\lvert x_i - \bar{x}\rvert = 70.64$

Quartile Deviation:

Quartile deviation is one of the measures of dispersion. Before getting into a deeper understanding, let's recall quartiles and how we can define them. Quartiles are the values that divide a list of numerical data into three-quarters, such as Q_1, Q_2 and Q_3. The middle part of the three quarters measures the central point of distribution and shows the data values near the midpoint (or the central value; this is referred to as the median). The lower part of the quarters indicates just half the information set, which comes under the median, and the upper part shows the remaining half, which falls above the median. Thus, the quartiles represent the distribution or dispersion of the given data set.

Quartile Deviation in Statistics can be defined as the statistic that measures the dispersion. Here, the Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.

Quartile Deviation Definition:

The Quartile Deviation can be defined mathematically as half of the difference between the upper and lower quartile. Here, quartile deviation can be represented as QD; Q_3 denotes the upper quartile and Q_1 indicates the lower quartile.

Quartile Deviation is also known as the Semi Interquartile range.

Quartile Deviation Formula:

Suppose Q_1 is the lower quartile, Q_2 is the median, and Q_3 is the upper quartile for the given data set, then its quartile deviation can be calculated using the following formula.

$QD = (Q_3 – Q_1)/_2$

In the next section, you will learn how to calculate these quartiles for both ungrouped and grouped data separately.

Quartile Deviation for Ungrouped Data:

For an ungrouped data, quartiles can be obtained using the following formulas,

Q_1 = [(n+1)/4]th item

Q_2 = [(n+1)/2]th item

Q_3 = [3(n+1)/4]th item

Where n represents the total number of observations in the given data set.

Also, Q_2 is the median of the given data set, Q_1 is the median of the lower half of the data set and Q_3 is the median of the upper half of the data set.

Before, estimating the quartiles, we have to arrange the given data values in ascending order. If the value of n is even, we can follow the similar procedure of finding the median.

Click here to get the quartiles calculator for ungrouped data.

Quartile Deviation for Grouped Data

For a grouped data, we can find the quartiles using the formula,

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f}(l_{2-}l_1)$$

Here,

Q_r = the rth quartile

l_1 = the lower limit of the quartile class

l_2 = the upper limit of the quartile class

f = the frequency of the quartile class

c = the cumulative frequency of the class preceding the quartile class

N = Number of observations in the given data set

Quartile Deviation Example:

Let's understand the quartile deviation of ungrouped and grouped data with the help of examples given below.

Example 1:

Find the quartiles and quartile deviation of the following data:

17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Solution:

Given data:

17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Ascending order of the given data is:

2, 5, 7, 7, 8, 8, 10, 10, 14, 15, 17, 18, 24, 27, 28, 48

Number of data values = n = 16

Q_2 = Median of the given data set

n is even, median = (1/2) [(n/2)th observation and (n/2 + 1)th observation]

= (1/2)[8th observation + 9th observation]

= (10 + 14)/2

= 24/2

= 12

Q_2 = 12

Now, lower half of the data is:

2, 5, 7, 7, 8, 8, 10, 10 (even number of observations)

Q_1 = Median of lower half of the data

= (1/2)[4th observation + 5th observation]

= (7 + 8)/2

= 15/2

= 7.5

Also, the upper half of the data is:

14, 15, 17, 18, 24, 27, 28, 48 (even number of observations)

Q_3= Median of upper half of the data

= (1/2)[4th observation + 5th observation]

= (18 + 24)/2

= 42/2

= 21

Quartile deviation = $(Q_3 – Q_1)/2$

= (21 – 7.5)/2

= 13.5/2

= 6.75

Therefore, the quartile deviation for the given data set is 6.75.

Example 2:

Calculate the quartile deviation for the following distribution.

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	5	3	4	3	3	4	7	9	7	8

Solution:

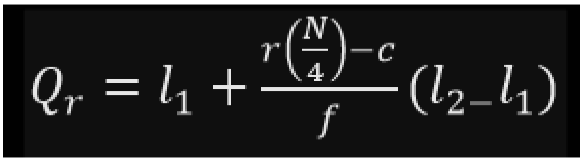
Let us calculate the cumulative frequency for the given distribution of data.

Class	Frequency	Cumulative Frequency
0 – 10	5	5
10 – 20	3	5 + 3 = 8
20 – 30	4	8 + 4 = 12
30 – 40	3	12 + 3 = 15
40 – 50	3	15 + 3 = 18
50 – 60	4	18 + 4 = 22
60 – 70	7	22 + 7 = 29
70 – 80	9	29 + 9 = 38

80 – 90	7	38 + 7 = 45
90 – 100	8	45 + 8 = 53

Here, N = 53

We know that,



$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right)-c}{f}(l_{2-}l_1)$$

Finding Q₁:

r = 1

N/4 = 53/4 = 13.25

Thus, Q1 lies in the interval 30 – 40.

In this case, quartile class = 30 – 40

l_1 = the lower limit of the quartile class = 30

l_2 = the upper limit of the quartile class = 40

f = the frequency of the quartile class = 3

c = the cumulative frequency of the class preceding the quartile class = 12

Now, by substituting these values in the formula we get:

Q_1 = 30 + [(13.25 – 12)/3] × (40 – 30)

= 30 + (1.25/3) × 10

= 30 + (12.5/3)

= 30 + 4.167

= 34.167

Finding Q_3:

r = 3

3N/4 = 3 × 13.25 = 39.75

Thus, Q_3 lies in the interval 80 – 90.

In this case, quartile class = 80 – 90

l_1 = the lower limit of the quartile class = 80

l_2 = the upper limit of the quartile class = 90

f = the frequency of the quartile class = 7

c = the cumulative frequency of the class preceding the quartile class = 38

Now, by substituting these values in the formula we get:

Q_3 = 80 + [(39.75 – 38)/7] × (90 – 80)

= 80 + (1.75/7) × 10

= 80 + (17.5/7)

= 80 + 2.5

= 82.5

Finally, the quartile deviation = (Q_3 – Q_1)/2

QD = (82.5 – 34.167)/2

= 48.333/2

= 24.1665

Hence, the quartile deviation of the given distribution is 24.167 (approximately).

Skewness

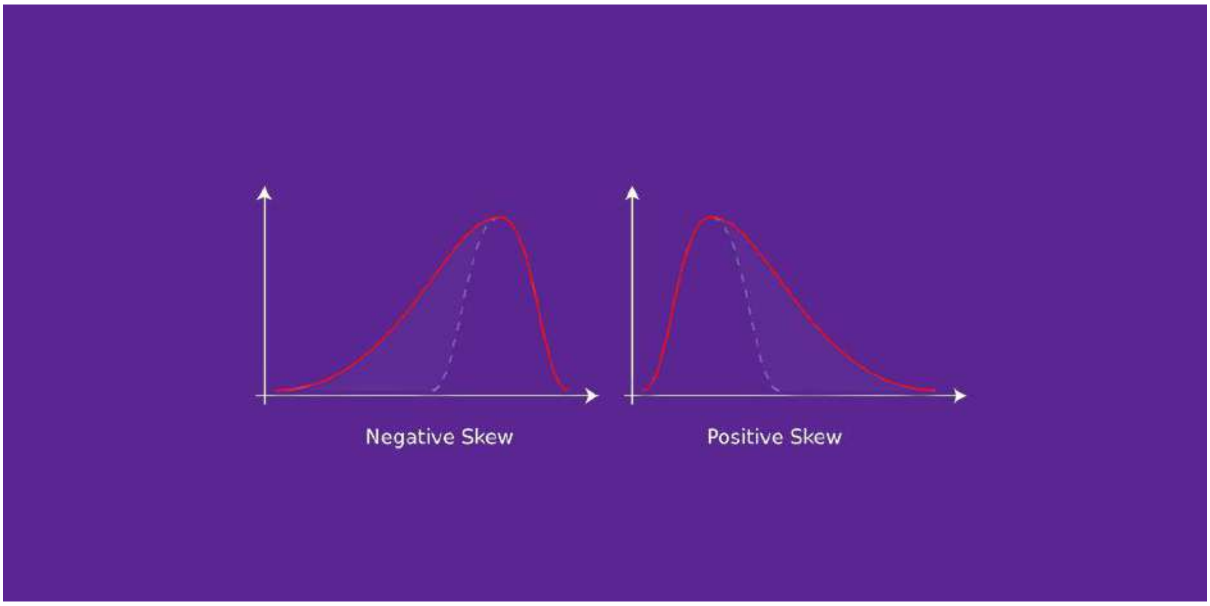
- Skewness is a key statistics concept you must know in the data science and analytics fields
- Learn what is skewness, and why it’s important for you as a data science professional

Introduction:

The concept of skewness is baked into our way of thinking. When we look at a visualization, our minds intuitively discern the pattern in that chart.

As you might already know, India has more than 50% of its population below the age of 25 and more than 65% below the age of 35. If you’ll plot the distribution of the age of the population of India, you will find that there is a hump on the left side of distribution and the right side is comparatively planar. In other words, we can say that there’s a skew towards the end, right?

So even if you haven’t read up on skewness as a data science or analytics professional, you have definitely interacted with the concept on an informal note. And it’s actually a pretty easy topic in statistics – and yet a lot of folks skim through it in their haste of learning other seemingly complex data science concepts. To me, that’s a mistake.



Skewness is a fundamental statistics concept that everyone in data science and analytics needs to know. It is something that we simply can't run away from. And I'm sure you'll understand this by the end of this article.

Here, we'll be discussing the concept of skewness in the easiest way possible. You'll learn about skewness, its types, and its importance in the field of data science. So buckle up because you'll learn a concept that you'll value during your entire data science career.

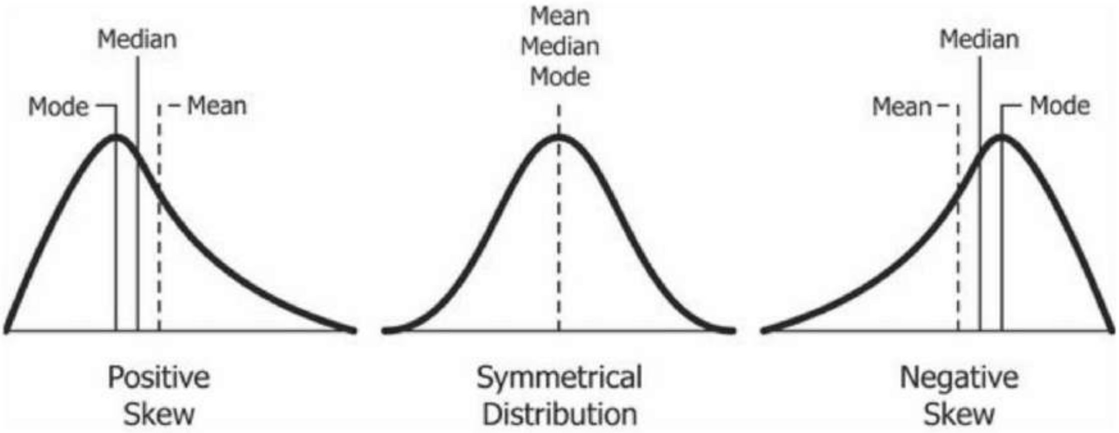
What is Skewness?

Skewness is the measure of the asymmetry of an ideally symmetric probability distribution and is given by the third standardized moment. If that sounds way too complex, don't worry! Let me break it down for you.

In simple words, skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution. Now, you might be thinking – why am I talking about normal distribution here?

Well, the normal distribution is the probability distribution without any skewness. You can look at the image below which shows symmetrical distribution that's basically a normal distribution and you can see that it is symmetrical on both sides of the dashed line. Apart from this, there are two types of skewness:

- Positive Skewness
- Negative Skewness



The probability distribution with its tail on the right side is a positively skewed distribution and the one with its tail on the left side is a negatively skewed distribution. If you're finding the above figures confusing, that's alright. We'll understand this in more detail later.

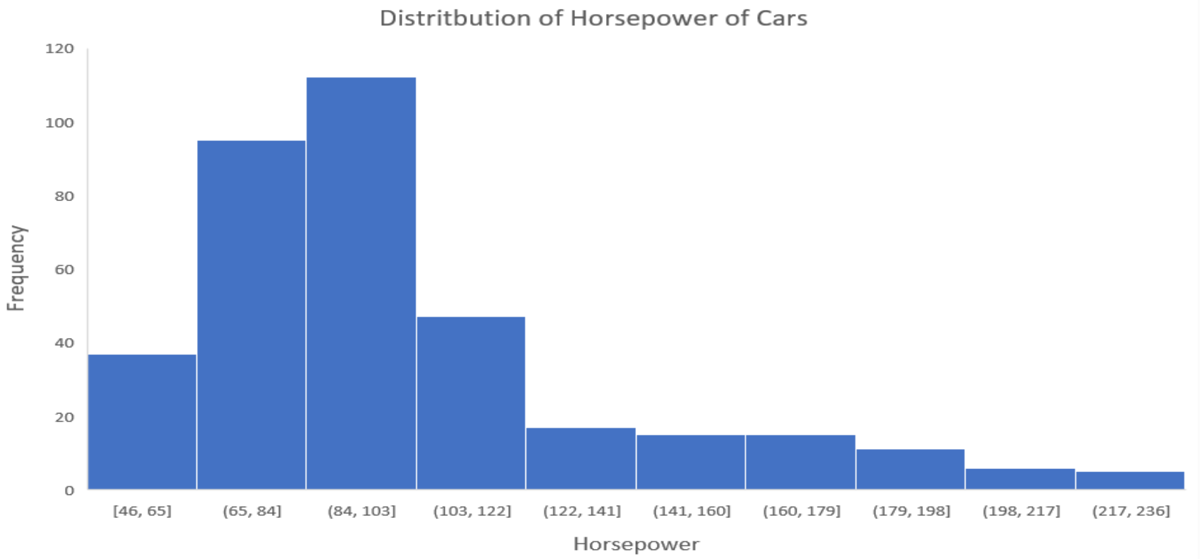
Before that, let's understand why skewness is such an important concept for you as a data science professional.

Why is Skewness Important?

Now, we know that the skewness is the measure of asymmetry and its types are distinguished by the side on which the tail of probability distribution lies. But why is knowing the skewness of the data important?

First, linear models work on the assumption that the distribution of the independent variable and the target variable are similar. Therefore, knowing about the skewness of data helps us in creating better linear models.

Secondly, let's take a look at the below distribution. It is the distribution of horsepower of cars:



You can clearly see that the above distribution is positively skewed. Now, let's say you want to use this as a feature for the model which will predict the mpg (miles per gallon) of a car.

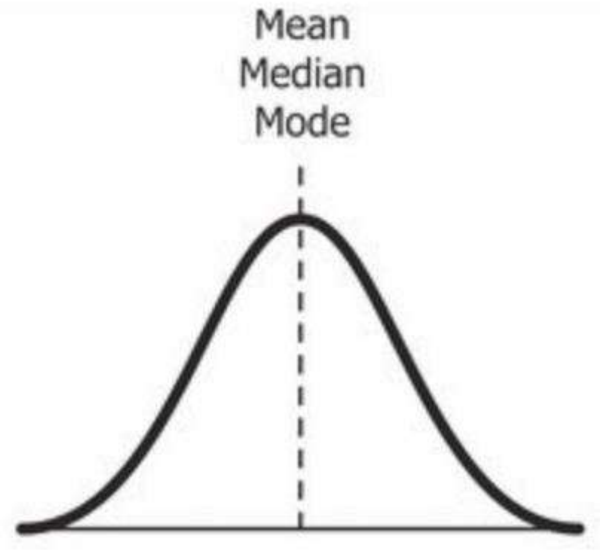
Since our data is positively skewed here, it means that it has a higher number of data points having low values, i.e., cars with less horsepower. So when we train our model on this data, it will perform better at predicting the mpg of cars with lower horsepower as compared to those with higher horsepower.

Also, skewness tells us about the direction of outliers. You can see that our distribution is positively skewed and most of the outliers are present on the right side of the distribution.

Note: *The skewness does not tell us about the number of outliers. It only tells us the direction.*

Now we know why skewness is important, let's understand the distributions which I showed you earlier.

What is Symmetric/Normal Distribution?



Credits: Wikipedia

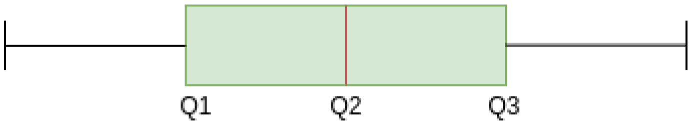
Yes, we're back again with the normal distribution. It is used as a reference for determining the skewness of a distribution. As I mentioned earlier, the ideal normal distribution is the probability distribution with almost no skewness. It is nearly perfectly symmetrical. Due to this, the value of skewness for a normal distribution is zero.

But, why is it nearly perfectly symmetrical and not absolutely symmetrical?

That’s because, in reality, no real word data has a perfectly normal distribution. Therefore, **even the value of skewness is not exactly zero; it is nearly zero.** Although the value of zero is used as a reference for determiningthe skewness of a distribution.

You can see in the above image that the same line represents the mean, median, and mode. It is because the mean, median, and mode of a perfectly normal distribution are equal.

So far, we’ve understood the skewness of normal distribution using a probability or frequency distribution. Now, let’s understand it in terms of a boxplot because that’s the most common way of looking at a distribution in the data science space.



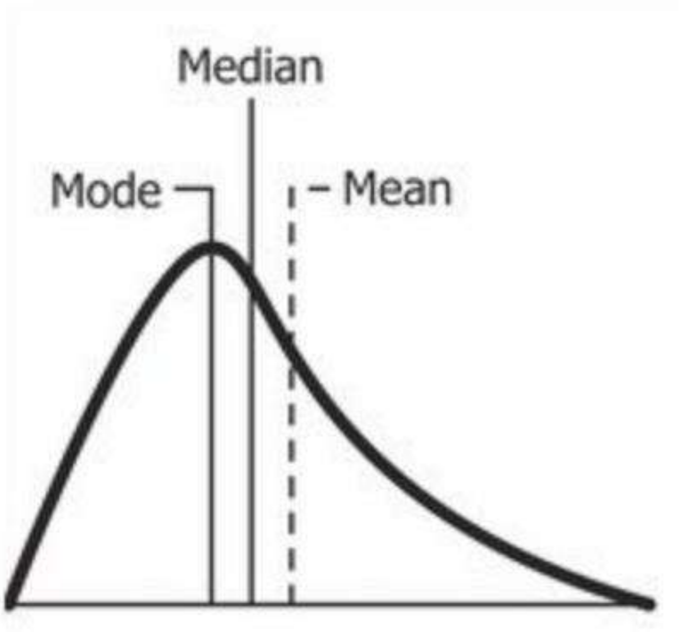
The above image is a boxplot of symmetric distribution. You’ll notice here that the distance between Q1 and Q2 and Q2 and Q3 is equal i.e.:

$$Q_3 - Q_2 = Q_2 - Q_1$$

But that’s not enough for concluding if a distribution is skewed or not. We also take a look at the length of the whisker; if they are equal, then we can say that the distribution is symmetric, i.e. it is not skewed.

Now that we’ve discussed the skewness in the normal distribution, it’s time to learn about the two types of skewness which we discussed earlier. Let’s start with positive skewness.

Understanding Positively Skewed Distribution

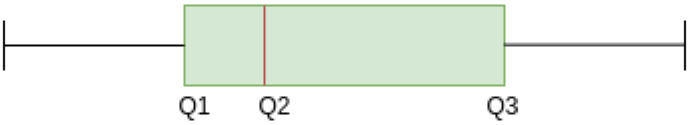


Source: Wikipedia

A positively skewed distribution is the distribution with the tail on its right side. The value of skewness for a positively skewed distribution is greater than zero. As you might have already understood by looking at the figure, the value of mean is the greatest one followed by median and then by mode.

So why is this happening?

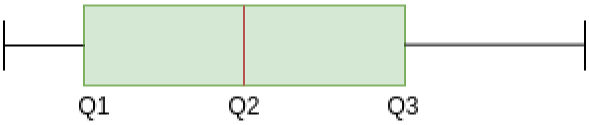
Well, the answer to that is that the skewness of the distribution is on the right; it causes the mean to be greater than the median and eventually move to the right. Also, the mode occurs at the highest frequency of the distribution which is on the left side of the median. Therefore, **mode < median < mean**.



In the above boxplot, you can see that Q2 is present nearer to Q1. This represents a positively skewed distribution. In terms of quartiles, it can be given by:

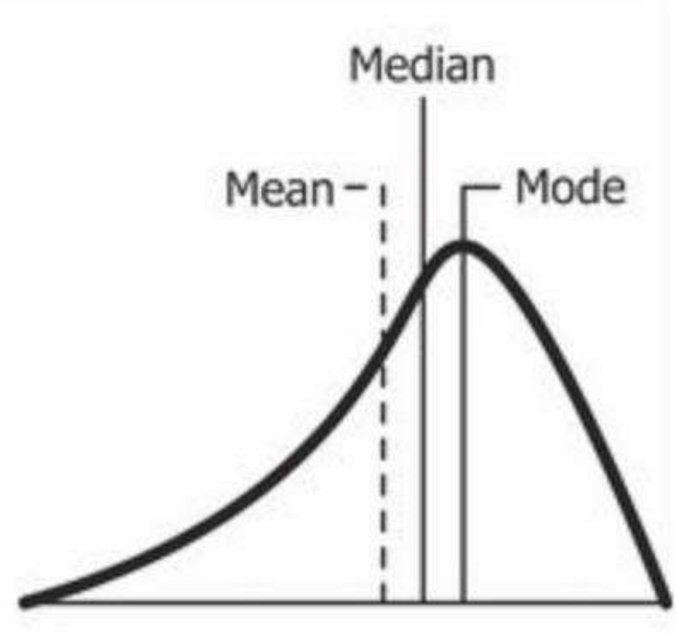
$$Q_3 - Q_2 > Q_2 - Q_1$$

In this case, it was very easy to tell if the data is skewed or not. But what if we have something like this:



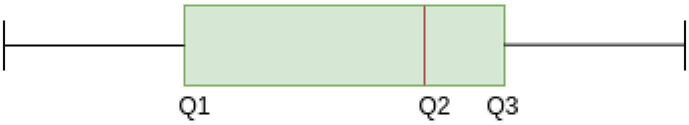
Here, Q2-Q1 and Q3-Q2 are equal and yet the distribution is positively skewed. The keen-eyed among you will have noticed the length of the right whisker is greater than the left whisker. From this, we can conclude that the data is positively skewed.

Understanding Negatively Skewed Distribution



Source: Wikipedia

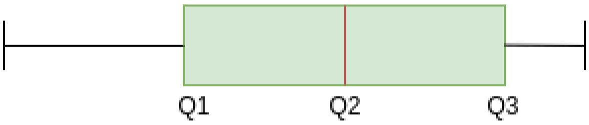
As you might have already guessed, a negatively skewed distribution is the distribution with the tail on its left side. The value of skewness for a negatively skewed distribution is less than zero. You can also see in the above figure that the **mean < median < mode**.



In the boxplot, the relationship between quartiles for a negative skewness is given by:

$$Q_3 - Q_2 < Q_2 - Q_1$$

Similar to what we did earlier, if Q3-Q2 and Q2-Q1 are equal, then we look for the length of whiskers. And if the length of the left whisker is greater than that of the right whisker, then we can say that the data is negatively skewed.



Skewness Formula

Skewness formula is called so because the graph plotted is displayed in skewed manner. Skewness is a measure used in statistics that helps reveal the asymmetry of a probability distribution. It can either be positive or negative, irrespective of signs. To calculate the skewness, we have to first find the mean and variance of the given data.

The formula is:

The formula is:

$$g = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^3}$$

Where,

x¯ is the sample mean
xi is the ith sample
n is the total number of observations
s is the standard deviation
g= sample skewness

Solved example

Question. Find the skewness in the following data.

Height (inches)	Class Marks	Frequency
59.5 – 62.5	61	5
62.5 – 65.5	64	18
65.5 – 68.5	67	42
68.5 – 71.5	70	27
71.5 – 74.5	73	8

To know how skewed these data are as compared to other data sets, we have to compute the skewness.

Sample size and sample mean should be found out.

N = 5 + 18 + 42 + 27 + 8 = 100

$$\bar{x} = \frac{(61 \times 5) + (64 \times 18) + (67 \times 42) + (70 \times 27) + (73 \times 8)}{100}$$

$$\bar{x} = \frac{6745}{100} = 67.45$$

Now with the mean we can compute the skewness.

Class Mark, *x*	Frequency, *f*	*xf*	$(x-\bar{x})$	$(x-\bar{x})^2 \times f$	$(x-\bar{x})^3 \times f$
61	5	305	-6.45	208.01	-1341.68
64	18	1152	-3.45	214.25	-739.15
67	42	2814	-0.45	8.51	-3.83
70	27	1890	2.55	175.57	447.70
73	8	584	5.55	246.42	1367.63
		6745	n/a	852.75	-269.33
		67.45	n/a	8.5275	-2.6933

Now, the skewness is

$$g = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^3}$$

s=√[(8.5275/(100-1))=0.2935]

g=√[(-2.693/[99 * (0.295)^3] = -1.038

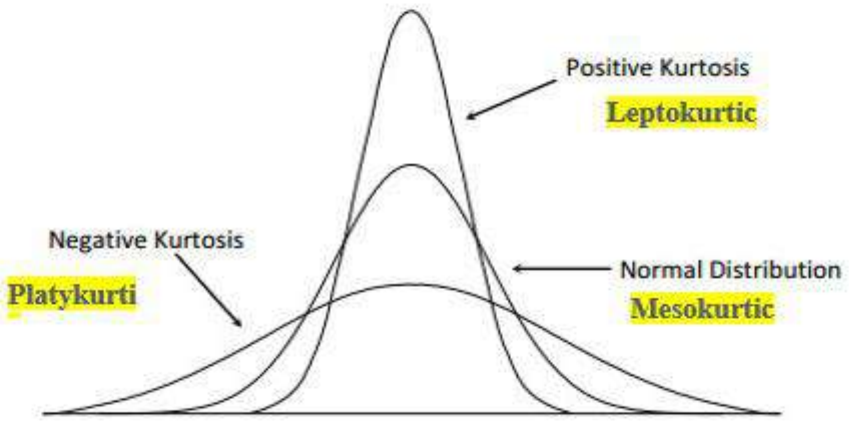
For interpreting we have the folowing rules as per Bulmer in the year 1979:

- If the skewness comes to less than -1 or greater than +1, the data distribution is highly skewed
- If the skewness comes to between -1 and −12 or between +12 and +1, the data distribution is moderately skewed.
- If the skewness is between −12 and +12,the distribution is approximately symmetric.

Kurtosis:

Kurtosis refers to the degree of presence of outliers in the distribution.

Kurtosis is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.



In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

Excess Kurtosis

The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3.

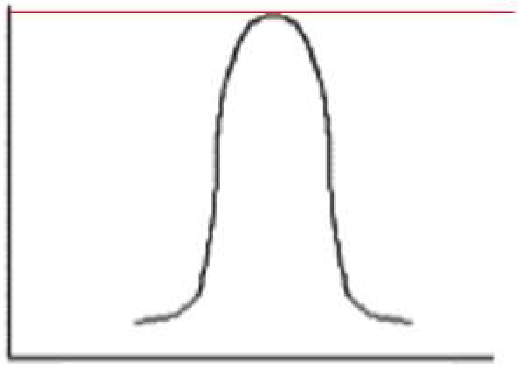
Excess kurtosis = Kurt – 3

Types of excess kurtosis

1. *Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).*
2. *Mesokurtic (kurtosis same as the normal distribution).*
3. *Platykurtic or short-tailed distribution (kurtosis less than normal distribution).*

Leptokurtic (kurtosis > 3)

Leptokurtic is having very long and skinny tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.



Leptokurtic

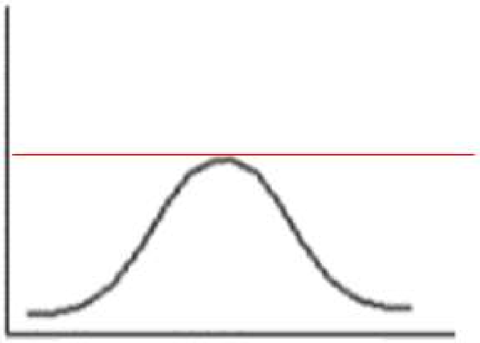
platykurtic (kurtosis < 3)

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

Mesokurtic (kurtosis = 3)

Mesokurtic is the same as the normal distribution, which means kurtosis is near to 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

Mesokurtic = 3 – 3 = 0



Mesokurtic

Summary:

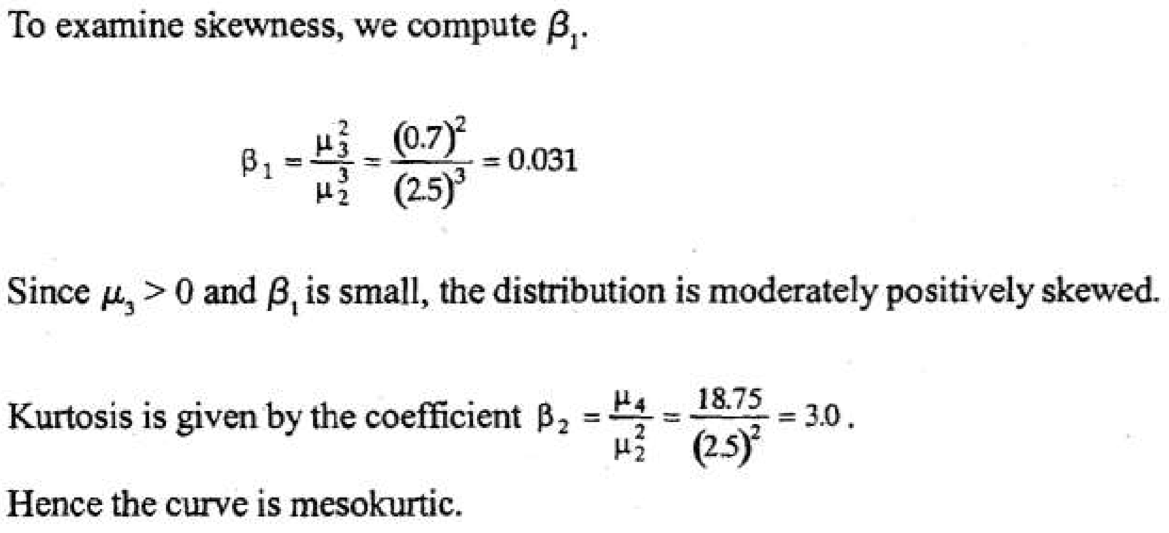
The skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution. Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).

When data skewed, the tail region may behave as an outlier for the statistical model, and outliers unsympathetically affect the model’s performance especially regression-based models. Some statistical models are hardy to outliers like Tree-based models, but it will limit the possibility to try other models. So there is a necessity to transform the skewed data to close enough to a Normal distribution.

Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). *Leptokurtic distribution (kurtosis more than normal distribution).Mesokurtic distribution (kurtosis same as the normal distribution).Platykurtic distribution (kurtosis less than normal distribution).*

A measure of kurtosis is given by $\beta_2 = \frac{\mu_4}{\mu_2^2}$, a coefficient given by Karl Pearson.The value of $\beta_2 = 3$ for a mesokurtic curve. When $\beta_2 > 3$, the curvt: is more peaked than the mesokurtic curve and is tenned as leptokurtic. Similarly, when $\beta_2 < 3$, the curve is less peaked than the mesokurtic curve and is called as platykurtic curve.

Example: The first four central moments of a distribution are 0,2.5,0.7 and 18.75. Examine the skewness and kurtosis of the distribution.



To examine skewness, we compute β_1.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.031$$

Since $\mu_3 > 0$ and β_1 is small, the distribution is moderately positively skewed.

Kurtosis is given by the coefficient $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3.0$.

Hence the curve is mesokurtic.

Lorenz Curve:

A Lorenz curve is a graphical representation of income inequality or wealth inequality developed by American economist Max Lorenz in 1905. The graph plots percentiles of the population on the horizontal axis according to income or wealth. It plots cumulative income or wealth on the vertical axis, so that an x-value of 45 and a y-value of 14.2 would mean that the bottom 45% of the population controls 14.2% of the total income or wealth. In practice, a Lorenz curve is usually a mathematical function estimated from an incomplete set of observations of income or wealth.

Understanding the Lorenz Curve:

The Lorenz curve is often accompanied by a straight diagonal line with a slope of 1, which represents perfect equality in income or wealth distribution; the Lorenz curve lies beneath it, showing the observed or estimated distribution. The area between the straight line and the curved line, expressed as a ratio of the area under the straight line, is the Gini coefficient, a scalar measurement of inequality.

While the Lorenz curve is most often used to represent economic inequality, it can also demonstrate unequal distribution in any system. The farther away the curve is from the baseline, represented by the straight diagonal line, the higher the level of inequality. In economics, the Lorenz curve denotes inequality in the distribution of either wealth or income; these are not synonymous since it is possible to have high earnings but zero or negative net worth, or low earnings but a large net worth.

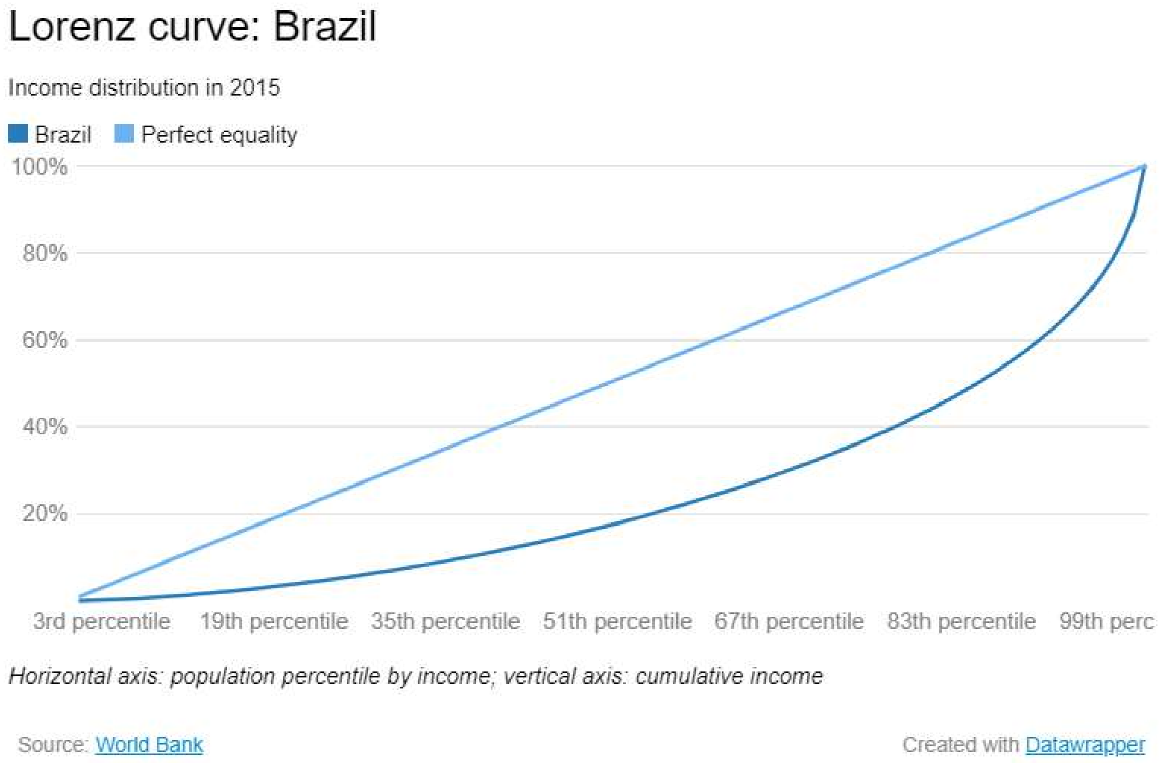
A Lorenz curve usually starts with an empirical measurement of wealth or income distribution across a population based on data such as tax returns which report income for a large portion of the population. A graph of the data may be used directly as a Lorenz curve, or economists and statisticians may fit a curve that represents a continuous function to fill in any gaps in the observed data.

A Lorenz curve gives more detailed information about the exact distribution of wealth or income across a population than summary statistics such as the Gini coefficient or the Lorenz asymmetry coefficient. Because a Lorenz curve visually displays the distribution across each percentile (or other unit breakdown), it can show precisely at which income (or wealth) percentiles the observed distribution varies from the line of equality and by how much.

However, because constructing a Lorenz curve involves fitting a continuous function to some incomplete set of data, there is no guarantee that the values along a Lorenz curve (other than those actually observed in the data) actually correspond to the true distributions of income. Most of the points along the curve are just guesses based on the shape of the curve that best fits the observed data points. So the shape of the Lorenz curve can be sensitive to the quality and sample size of the data and to the mathematical assumptions and judgments as to what constitutes a best fit curve, and these may represent sources of substantial error between the Lorenz curve and the actual distribution.

Lorenz Curve Example:

The Gini coefficient is used to express the extent of inequality in a single figure. It can range from 0 (or 0%) to 1 (or 100%). Complete equality, in which every individual has the exact same income or wealth, corresponds to a coefficient of 0. Plotted as a Lorenz curve, complete equality would be a straight diagonal line with a slope of 1 (the area between this curve and itself is 0, so the Gini coefficient is 0). A coefficient of 1 means that one person earns all of the income or holds all of the wealth. Accounting for negative wealth or income, the figure can theoretically be higher than 1; in that case, the Lorenz curve would dip below the horizontal axis.


Lorenz curve: Brazil

Income distribution in 2015

Horizontal axis: population percentile by income; vertical axis: cumulative income

Source: World Bank

Created with Datawrapper

The curve above shows a continuous Lorenz curve that has been fitted to the data that describe the income distribution in Brazil in 2015, compared to a straight diagonal line representing perfect equality. At the 55th income percentile, the value of the Lorenz curve is 20.59%: in other words, this Lorenz curve estimates that the bottom 55% of the population takes in 20.59% of the nation's total income. If Brazil were a perfectly equal society, the bottom 55% would earn 55% of the total. The 99th percentile corresponds to 88.79% in cumulative income, meaning that the top 1% takes in 11.21% of Brazil's income.

To find the approximate Gini coefficent, subtract the area beneath the Lorenz curve (around 0.25) from the area beneath the line of perfect equality (0.5 by definition). Divide the result by the area beneath the line of perfect equality, which yields a coefficient of around 0.5 or 50%. According to the CIA, Brazil's Gini coefficient in 2014 was 49.7%.

UNIT – III

Correlation and Regression Analysis

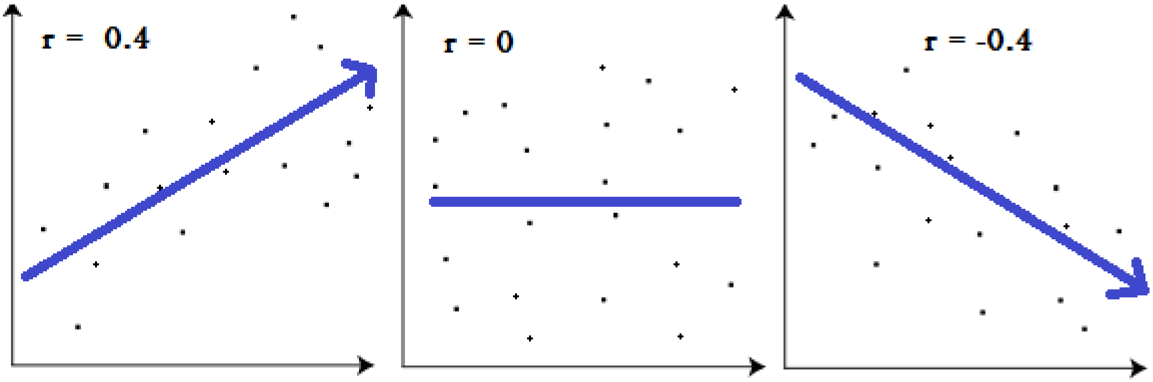
Correlation Analysis

Correlation is a term that is a measure of the strength of a linear relationship between two quantitative variables (e.g., height, weight).

Positive correlation is a relationship between two variables in which both variables move in the same direction. This is when one variable increases while the other increases and visa versa. For example, *positive* correlation may be that the more you exercise, the more calories you will burn. Whilst *negative correlation* is a relationship where one variable increases as the other decreases, and vice versa.

Correlation coefficients are used to measure how strong a relationship is between two variables
Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



Positive Correlation **No correlation** **Negative**

scatter diagram:

The scatter diagram is a technique used to examine the relationship between both the axis (X and Y) with one variable. In the graph, if the variables are correlated, the point will drop along a curve or line. A scatter diagram or scatter plot, is used to give an idea idea of the nature of relationship.

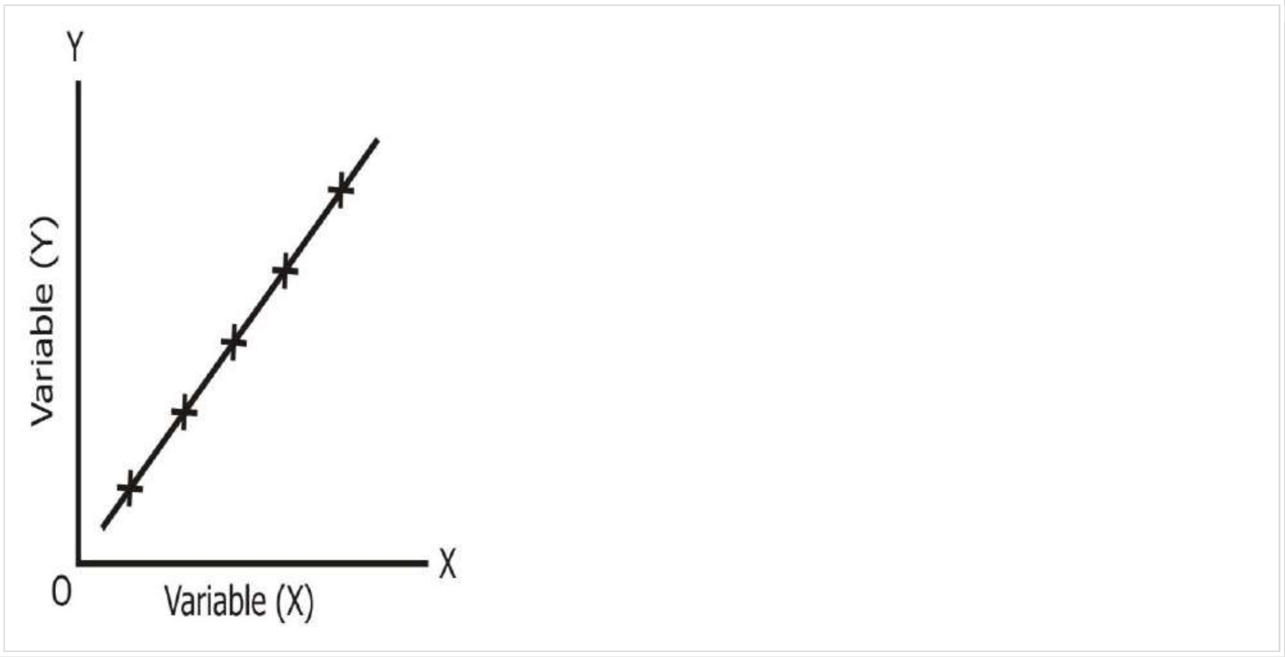
In scatter correlation diagram, if all the points stretches in one line, then the correlation is perfect and is in unity. But,if the scatter points are widely scattered throughout the line, the correlation is said to be low. And if the scatter points rest near a line or on a line the correlation is said to be linear.

EX.1 FROM THE FOLLOWING DRAW A SCATTER DIAGRAM AND STATE THE TYPE OF CORRELATION BETWEEN THE VARIABLE X AND Y:

X	1	2	3	4	5
Y	5	10	15	20	25

Answer:

Perfect Positive Correlation



The Karl Pearson’s correlation coefficient:

A correlation coefficient is generally applied in statistics to calculate a relationship between two variables. The correlation shows a specific value of a degree of a linear relationship between X and Y variables. There are various types of correlation coefficient. However, Pearson’s correlation (also called Pearson’s R) is the correlation coefficient frequently used in linear regression.

Pearson’s Coefficient Correlation

Karl Pearson’s Coefficient of Correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linear related variables. The coefficient of correlation is expressed by **“r”.**

Karl Pearson Correlation Coefficient Formula:

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

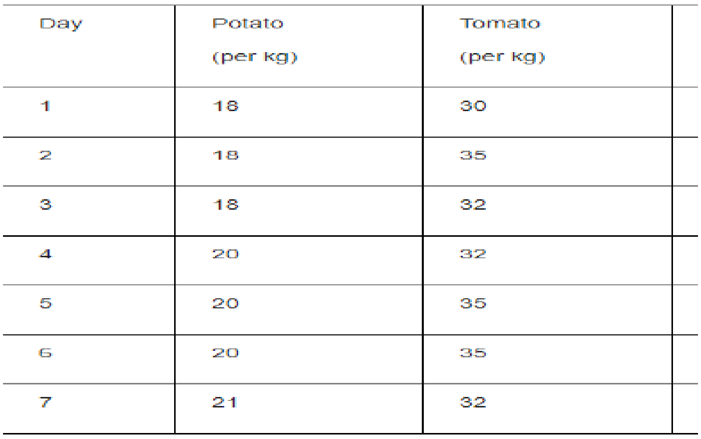
Alternative Formula (covariance formula):

$$\mathrm{Cov}(X,Y) = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{N} = \frac{\Sigma xy}{N}$$

Pearson correlation example

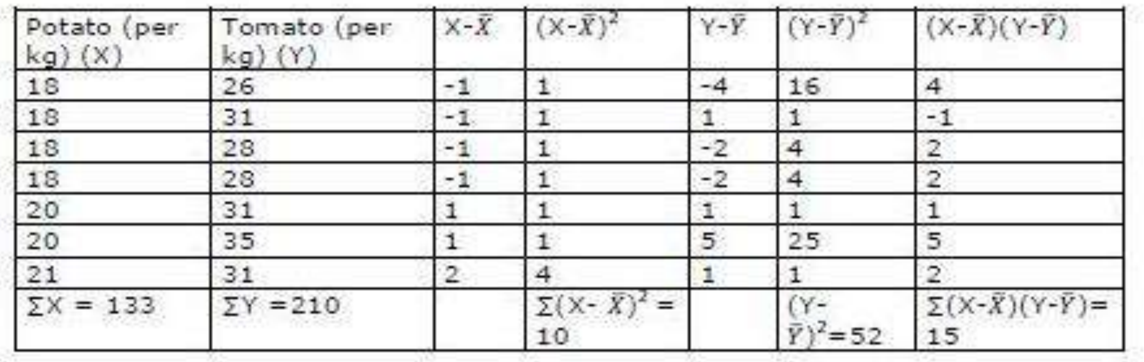
- When a correlation coefficient is (1) that means every increase in one variable, there is a positive increase in other fixed proportion. For instance, shoe sizes change according to the length of the foot and are (almost) perfect correlation.
- When a correlation coefficient is (-1) that means every positive increase in one variable, there is a negative decrease in other fixed proportion. For instance, with the decrease in the quantity of gas in a gas tank, it shows (almost) a perfect correlation with speed.
- When a correlation coefficient is (0) for every increase, it means there is no positive or negative increase, and the two variables are not related.

Example Problem 2:



Day	Potato (per kg)	Tomato (per kg)
1	18	30
2	18	35
3	18	32
4	20	32
5	20	35
6	20	35
7	21	32

Find the Karl Pearson's Correlation Coefficient between Potato and Tomato Price.

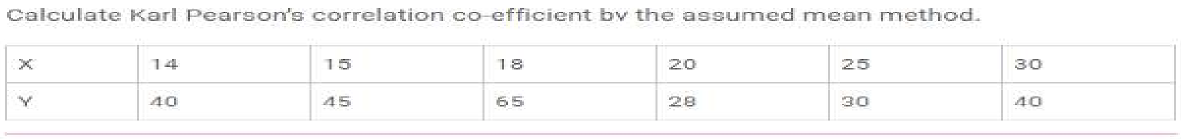


Potato (per kg) (X)	Tomato (per kg) (Y)	X-$\bar{X}$	(X-$\bar{X}$)^2	Y-$\bar{Y}$	(Y-$\bar{Y}$)^2	(X-$\bar{X}$)(Y-$\bar{Y}$)
18	26	-1	1	-4	16	4
18	31	-1	1	1	1	-1
18	28	-1	1	-2	4	2
18	28	-1	1	-2	4	2
20	31	1	1	1	1	1
20	35	1	1	5	25	5
21	31	2	4	1	1	2
ΣX = 133	ΣY =210		Σ(X- $\bar{X}$)^2 = 10		(Y-$\bar{Y}$)^2=52	Σ(X-$\bar{X}$)(Y-$\bar{Y}$)= 15

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})2(Y-\bar{Y})2}} = +0.65$$

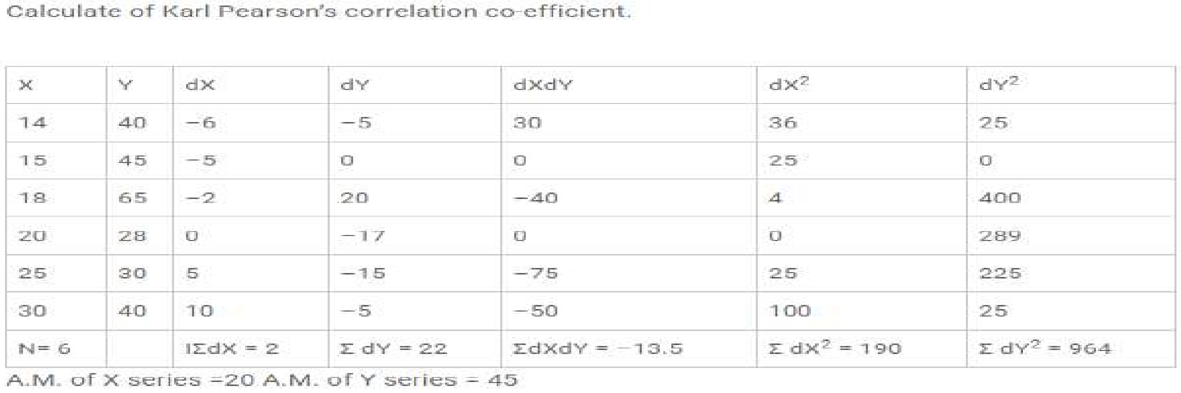
Karl Pearson's Assumed Mean Method:

Example Problem 3:



Calculate Karl Pearson's correlation co-efficient bv the assumed mean method.

X	14	15	18	20	25	30
Y	40	45	65	28	30	40



Calculate of Karl Pearson's correlation co-efficient.

X	Y	dX	dY	dXdY	dX^2	dY^2
14	40	−6	−5	30	36	25
15	45	−5	0	0	25	0
18	65	−2	20	−40	4	400
20	28	0	−17	0	0	289
25	30	5	−15	−75	25	225
30	40	10	−5	−50	100	25
N= 6		IΣdX = 2	Σ dY = 22	ΣdXdY = −13.5	Σ dX^2 = 190	Σ dY^2 = 964

A.M. of X series =20 A.M. of Y series = 45

$$r = \frac{\Sigma dXdY - \frac{\Sigma dX \Sigma dY}{N}}{\sqrt{\Sigma dX^2 - \frac{\Sigma dX^2}{N}}\sqrt{\Sigma dY^2 - \frac{\Sigma dY^2}{N}}}$$

$$\frac{-135 - \frac{2\times -22}{6}}{\sqrt{190 - \frac{(2)^2}{6}}.\sqrt{964 - \frac{(-22)^2}{6}}}$$

$$= \frac{-135 + \frac{44}{6}}{\sqrt{190 - \frac{4}{6}}.\sqrt{964 - \frac{484}{6}}} = \frac{\frac{-766}{6}}{\sqrt{\frac{568}{3}}.\sqrt{\frac{2650}{3}}} = -0.2 \quad \textbf{Ans.}$$

Spearman's rank-order correlation:

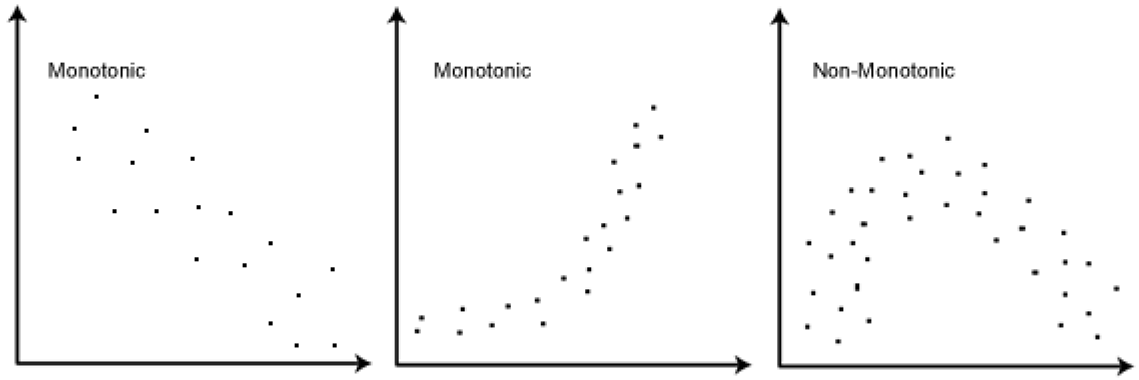
The Spearman's rank-order correlation is the nonparametric version of the **Pearson product-moment correlation**. Spearman's correlation coefficient, (ρ, also signified by r_s) measures the strength and direction of association between two ranked variables**.**

The assumptions of the test:

You need two variables that are either ordinal, interval or ratio (see our **Types of Variable** guide if you need clarification). Although you would normally hope to use a Pearson product-moment correlation on interval or ratio data, the Spearman correlation can be used when the assumptions of the Pearson correlation are markedly violated. However, Spearman's correlation determines the strength and direction of the monotonic relationship between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines.

What is a monotonic relationship?

A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases. Examples of monotonic and non-monotonic relationships are presented in the diagram below:



Why is a monotonic relationship important to Spearman's correlation?

Spearman's correlation measures the strength and direction of monotonic association between two variables. Monotonicity is "less restrictive" than that of a linear relationship. For example, the middle image above shows a relationship that is monotonic, but not linear.

A monotonic relationship is not strictly an assumption of Spearman's correlation. That is, you can run a Spearman's correlation on a non-monotonic relationship to determine if there is a **monotonic component** to the association. However, you would normally pick a measure of association, such as Spearman's correlation, that fits the pattern of the observed data. That is, if a scatterplot shows that the relationship between your two variables looks monotonic you would run a Spearman's correlation because this will then measure the strength and direction of this monotonic relationship. On the other hand if, for example, the relationship appears linear (assessed via scatterplot) you would run a Pearson's correlation because this will measure the strength and direction of any linear relationship. You will not always be able to visually check whether you have a monotonic relationship, so in this case, you might run a Spearman's correlation anyway.

How to rank data?

In some cases your data might already be ranked, but often you will find that you need to rank the data yourself. Thankfully, ranking data is not a difficult task and is easily achieved by working through your data in a table. Let us consider the following example data regarding the marks achieved in a maths and English exam:

	Marks									
English	56	75	45	71	61	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

The procedure for ranking these scores is as follows:

First, create a table with four columns and label them as below:

English (mark)	Maths (mark)	Rank (English)	Rank (maths)
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
61	65	6.5	5
64	56	5	9
58	59	8	8
80	77	1	1

English (mark)	Maths (mark)	Rank (English)	Rank (maths)
76	67	2	3
61	63	6.5	6

You need to rank the scores for maths and English separately. The score with the highest value should be labelled "1" and the lowest score should be labelled "10" (if your data set has more than 10 cases then the lowest score will be how many cases you have). Look carefully at the two individuals that scored 61 in the English exam (highlighted in bold). Notice their joint rank of 6.5. This is because when you have two identical values in the data (called a "tie"), you need to take the average of the ranks that they would have otherwise occupied. We do this because, in this example, we have no way of knowing which score should be put in rank 6 and which score should be ranked 7. Therefore, you will notice that the ranks of 6 and 7 do not exist for English. These two ranks have been averaged ((6 + 7)/2 = 6.5) and assigned to each of these "tied" scores.

What is the definition of Spearman's rank-order correlation?

There are two methods to calculate Spearman's correlation depending on whether: (1) your data does not have tied ranks or (2) your data has tied ranks. The formula for when there are no tied ranks is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i = difference in paired ranks and n = number of cases. The formula to use when there are tied ranks is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where i = paired score.

What values can the Spearman correlation coefficient, r_s, take?

The Spearman correlation coefficient, r_s, can take values from +1 to -1. A r_s of +1 indicates a perfect association of ranks, a r_s of zero indicates no association between ranks and a r_s of -1 indicates a perfect negative association of ranks. The closer r_s is to zero, the weaker the association between the ranks.

Example Problem 4:

An example of calculating Spearman's correlation:

To calculate a Spearman rank-order correlation on data without any ties we will use the following data:

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

We then complete the following table:

Where d = difference between ranks and d^2 = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

as n = 10. Hence, we have a ρ (or r_s) of 0.67. This indicates a strong positive relationship between the ranks individuals obtained in the maths and English exam. That is, the higher you ranked in maths, the higher you ranked in English also, and vice versa.

Repeated ranks

When two or more items have equal values (i.e., a tie) it is difficult to give ranks to them. In such cases the items are given the average of the ranks they would have received. For example, if two individuals are placed in the 8th place, they are given the rank [8+9] / 2 = 8.5 each, which is common rank to be assigned and the next will be 10; and if three ranked equal at the 8th place, they are given the rank [8 + 9 +10] /3 = 9 which is the common rank to be assigned to each; and the next rank will be 11.

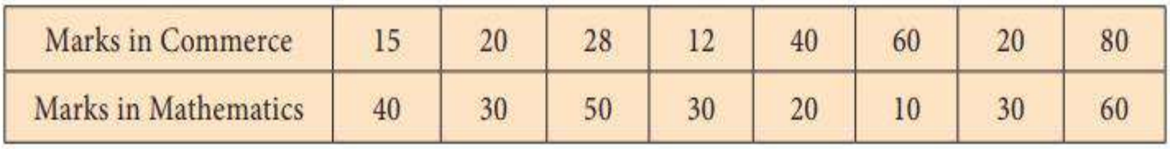
In this case, a different formula is used when there is more than one item having the same value.

$$\rho = 1 - 6\left[\frac{\sum D_i^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{n\left(n^2 - 1\right)}\right]$$

where m_i is the number of repetitions of i^{th} rank

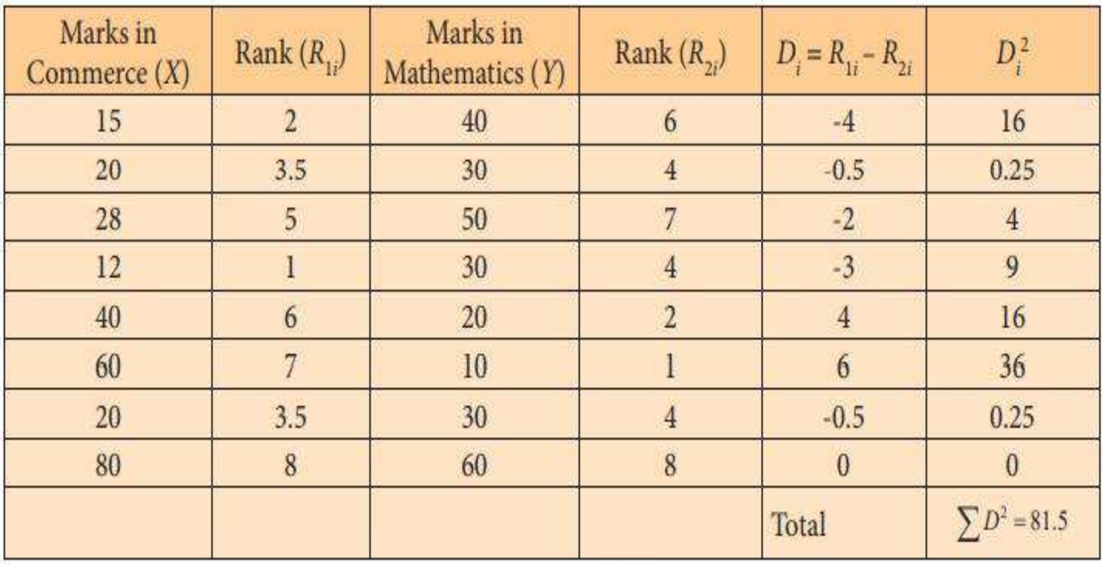
Example 5

Compute the rank correlation coefficient for the following data of the marks obtained by 8 students in the Commerce and Mathematics.



Marks in Commerce	15	20	28	12	40	60	20	80
Marks in Mathematics	40	30	50	30	20	10	30	60

Solution:



Marks in Commerce (X)	Rank (R_{1i})	Marks in Mathematics (Y)	Rank (R_{2i})	$D_i = R_{1i} - R_{2i}$	D_i^2
15	2	40	6	-4	16
20	3.5	30	4	-0.5	0.25
28	5	50	7	-2	4
12	1	30	4	-3	9
40	6	20	2	4	16
60	7	10	1	6	36
20	3.5	30	4	-0.5	0.25
80	8	60	8	0	0
				Total	$\sum D^2 = 81.5$

$$\rho = 1 - 6\left[\frac{\sum D_i^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{n\left(n^2 - 1\right)}\right]$$

Repetitions of ranks

In Commerce (X), 20 is repeated two times corresponding to ranks 3 and 4. Therefore, 3.5 is assigned for rank 2 and 3 with m_1=2.

In Mathematics (Y), 30 is repeated three times corresponding to ranks 3, 4 and 5. Therefore, 4 is assigned for ranks 3,4 and 5 with m_2=3.

Therefore,

$$\rho = 1 - 6\left[\frac{81.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)}{8\left(8^2 - 1\right)}\right]$$

$$= 1 - 6\frac{[81.5 + 0.5 + 2]}{504} = 1 - \frac{504}{504} = 0$$

Interpretation: Marks in Commerce and Mathematics are uncorrelated.

<u>RegressionAnalysis</u>:

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Regression helps investment and financial managers to value assets and understand the relationships between variables, such as commodity prices and the stocks of businesses dealing in those commodities.

Regression Explained
The two basic types of regression are simple linear regression and multiple linear regression, although there are non-linear regression methods for more complicated data and analysis. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome.

Regression can help finance and investment professionals as well as professionals in other businesses. Regression can also help predict sales for a company based on weather, previous sales, GDP growth, or other types of conditions. The capital asset pricing model (CAPM) is an often-used regression model in finance for pricing assets and discovering costs of capital.

The general form of each type of regression is:

- **Simple linear regression:** $Y = a + bX + u$
- **Multiple linear regression:** $Y = a + b_1X_1 {}^{+} b_2X_2 + b_3X_3 + ... + b_tX_t + u$

Where:

- Y = the variable that you are trying to predict (dependent variable).
- X = the variable that you are using to predict Y (independent variable).
- a = the intercept.
- b = the slope.
- u = the regression residual.

Regression takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a straight line (linear regression) that best approximates all the individual data points. In multiple regression, the separate variables are differentiated by using subscripts.

Linear Regression:

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, "how much additional sales income do I get for each additional $1000 spent on marketing?"

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, "what will the price of gold be in 6 months?"

If the plot of n pairs of data (x , y) for an experiment appear to indicate a "linear relationship" between y and x, then the method of least squares may be used to write a linear relationship between x and y.

The least squares regression line is the line that minimizes the sum of the squares (d1 + d2 + d3 + d4) of the vertical deviation from each data point to the line (see figure below as an example of 4 points).

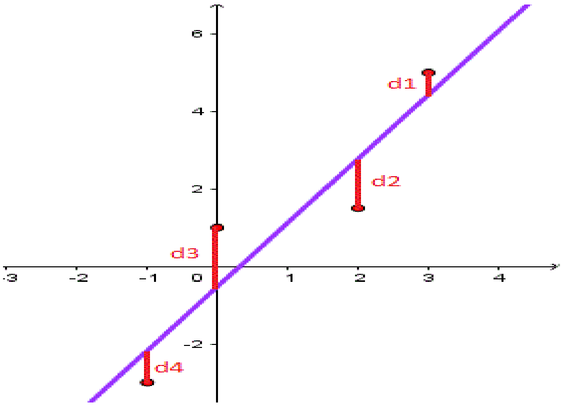


Figure 1. Linear regression where the sum of vertical distances d1 + d2 + d3 + d4 between observed and predicted (line and its equation) values is minimized.

The least square regression line for the set of n data points is given by the equation of a line in slope intercept form:

y = a x + b

where a and b are given by

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i)$$

Figure 2. Formulas for the constants a and b included in the linear regression .

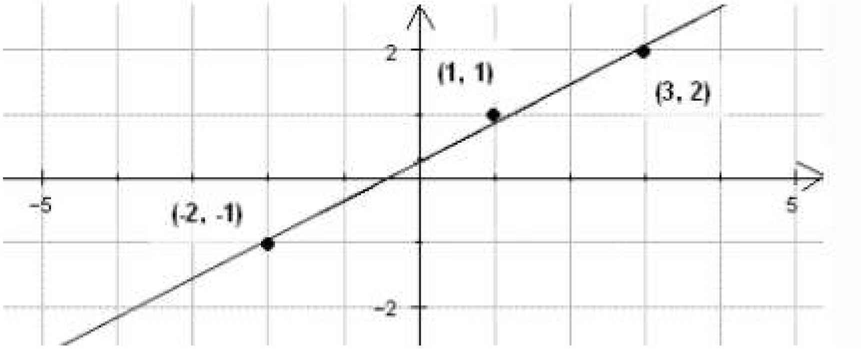
Problem 1

Consider the following set of points: {(-2 , -1) , (1 , 1) , (3 , 2)}
a) Find the least square regression line for the given data points.
b) Plot the given points and the regression line in the same rectangular system of axes.

1. a) Let us organize the data in a table.

x	**y**	**x y**	**x ^{2}**
-2	-1	2	4
1	1	1	1
3	2	6	9
Σx = 2	Σy = 2	Σxy = 9	Σx^2 = 14

We now use the above formula to calculate a and b as follows
a = (nΣx y - ΣxΣy) / (nΣx^2 - (Σx)^2) = (3*9 - 2*2) / (3*14 - 2^2) = 23/38
b = (1/n)(Σy - a Σx) = (1/3)(2 - (23/38)*2) = 5/19
b) We now graph the regression line given by y = a x + b and the given points.



Problem 2:

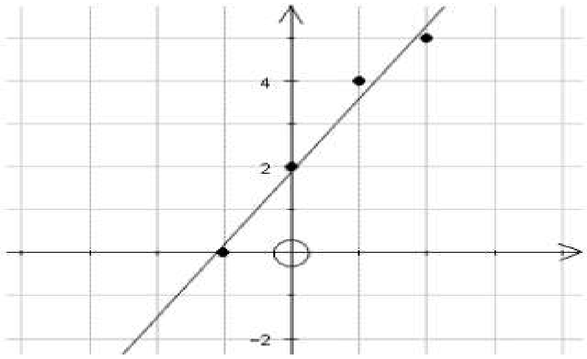
a) Find the least square regression line for the following set of data {(-1 , 0),(0 , 2),(1 , 4),(2 , 5)}

b) Plot the given points and the regression line in the same rectangular system of axes.

1. a) We use a table as follows

x	**y**	**x y**	**x ^{2}**
-1	0	0	1
0	2	0	0
1	4	4	1
2	5	10	4
$\Sigma x = 2$	$\Sigma y = 11$	$\Sigma x\ y = 14$	$\Sigma x^2 = 6$

We now use the above formula to calculate a and b as follows
$a = (n\Sigma x\ y - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2) = (4*14 - 2*11) / (4*6 - 2^2) = 17/10 = 1.7$
$b = (1/n)(\Sigma y - a\ \Sigma x) = (1/4)(11 - 1.7*2) = 1.9$
b) We now graph the regression line given by y = ax + b and the given points.



Problem 3

The values of y and their corresponding values of y are shown in the table below

X	0	1	2	3	4
Y	2	3	5	4	6

a) Find the least square regression line y = a x + b.
b) Estimate the value of y when x = 10.

1. a) We use a table to calculate a and b.

x	**y**	**x y**	**x^2**
0	2	0	0
1	3	3	1
2	5	10	4
3	4	12	9
4	6	24	16
Σx = 10	Σy = 20	Σx y = 49	$\Sigma x^2 = 30$

We now calculate a and b using the least square regression formulas for a and b.
$a = (n\Sigma x\ y - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2) = (5*49 - 10*20) / (5*30 - 10^2) = 0.9$
b = (1/n)(Σy - a Σx) = (1/5)(20 - 0.9*10) = 2.2
b) Now that we have the least square regression line y = 0.9 x + 2.2, substitute x by 10 to find the value of the corresponding y.
y = 0.9 * 10 + 2.2 = 11.2

Problem 4:

The sales of a company (in million dollars) for each year are shown in the table below.

x (year)	2005	2006	2007	2008	2009
y (sales)	12	19	29	37	45

a) Find the least square regression line y = a x + b.
b) Use the least squares regression line as a model to estimate the sales of the company in 2012.

a) We first change the variable x into t such that t = x - 2005 and therefore t represents the number of years after 2005. Using t instead of x makes the numbers smaller and therefore manageable. The table of values becomes.

t (years after 2005)	0	1	2	3	4
y (sales)	12	19	29	37	45

We now use the table to calculate a and b included in the least regression line formula.

t	**y**	**t y**	**t^2**
0	12	0	0

1	19	19	1
2	29	58	4
3	37	111	9
4	45	180	16
$\Sigma x = 10$	$\Sigma y = 142$	$\Sigma xy = 368$	$\Sigma x^2 = 30$

We now calculate a and b using the least square regression formulas for a and b.
$a = (n\Sigma t\ y - \Sigma t\Sigma y) / (n\Sigma t^2 - (\Sigma t)^2) = (5*368 - 10*142) / (5*30 - 10^2) = 8.4$
$b = (1/n)(\Sigma y - a\ \Sigma x) = (1/5)(142 - 8.4*10) = 11.6$
b) In 2012, t = 2012 - 2005 = 7
The estimated sales in 2012 are: y = 8.4 * 7 + 11.6 = 70.4 million dollars.

CHAPTER – 4

TIME SERIES

TIME SERIES Time series is set of data collected and arranged in accordance of time. According to Croxton and Cowdon,"A Time series consists of data arranged chronologically." Such data may be series of temperature of patients, series showing number of suicides in different months of year etc. The analysis of time series means separating out different components which influences values of series. The variations in the time series can be divided into two parts: long term variations and short term variations.Long term variations can be divided into two parts: Trend or Secular Trend and Cyclical variations. Short term variations can be divided into two parts: Seasonal variations and Irregular Variations.*Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for.*

Traditional methods of ***time series analysis*** are concerned with *decomposing* of a series into a *trend*, a *seasonal variation*, and other *irregular fluctuations*. Although this approach is not always the best but still useful (Kendall and Stuart, 1996).

The components, by which time series is composed of, are called the component of time series data. There are four basic components of the time series data described below.

Different Sources of Variation are:

1. **Seasonal effect (Seasonal Variation or Seasonal Fluctuations):**

Many of the *time series data* exhibits a ***seasonal variation*** which is the annual period, such as sales and temperature readings. This type of variation is easy to understand and can be easily measured or removed from the data to give deseasonalized data. Seasonal Fluctuations describes any regular variation (fluctuation) with a period of less than one year for example cost of various types of fruits and vegetables, clothes, unemployment figures, average daily rainfall, increase in the sale of tea in winter, increase in the sale of ice cream in summer, etc., all show seasonal variations. The changes which repeat themselves within a fixed period, are also called ***seasonal variations***, for example, traffic on roads in morning and evening hours, Sales at festivals like EID, etc., increase in the number of passengers at weekend, etc. Seasonal variations are caused by climate, social customs, religious activities, etc.

2. **Other Cyclic Changes (Cyclical Variation or Cyclic Fluctuations):**

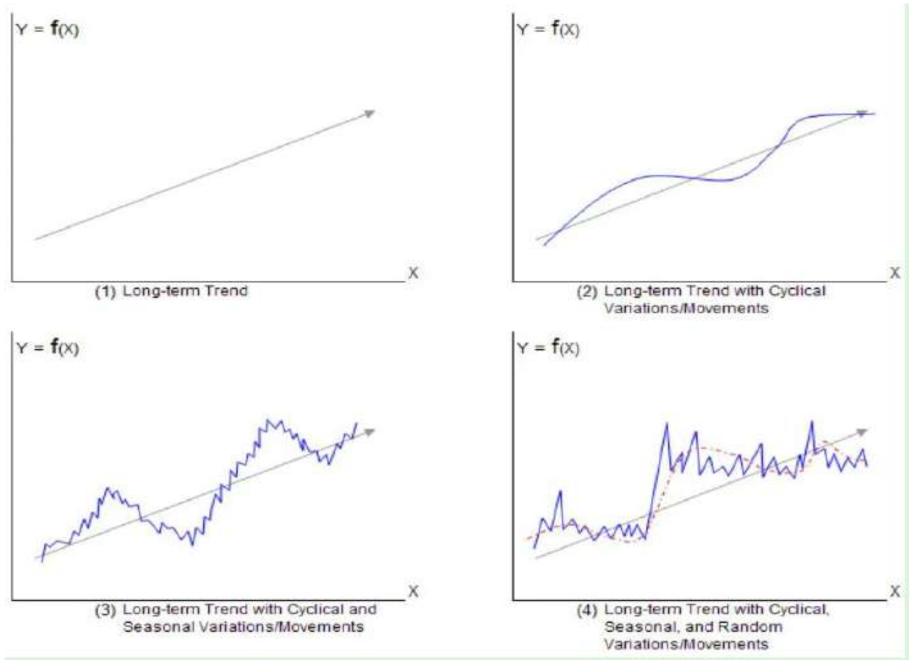
Time series exhibits ***Cyclical Variations*** at a fixed period due to some other physical cause, such as daily variation in temperature. ***Cyclical variation*** is a non-seasonal component that varies in a recognizable cycle. Sometimes series exhibits oscillation which does not have a fixed period but is predictable to some extent. For example, economic data affected by business cycles with a period varying between about 5 and 7 years. In weekly or monthly data, the cyclical component may describe any regular variation (fluctuations) in time series data. The cyclical variation is periodic in nature and repeats itself like a business cycle, which has four phases (i) *Peak* (ii) *Recession* (iii) *Trough/Depression* (iv) *Expansion*.

3. **Trend (Secular Trend or Long Term Variation):**

It is a longer-term change. Here we take into account the number of observations available and make a subjective assessment of what is long term. To understand the meaning of the long term, let for example climate variables sometimes exhibit cyclic variation over a very long time period such as 50 years. If one just had 20 years of data, this long term oscillation would appear to be a trend, but if several hundreds of years of data are available, then long term oscillations would be visible. These movements are systematic in nature where the movements are broad, steady, showing a slow rise or fall in the same direction. The trend may be linear or non-linear (curvilinear). Some examples of the secular trends are: Increase in prices, Increase in pollution, an increase in the need for wheat, increase in literacy rate, decrease in deaths due to advances in science. Taking averages over a certain period is a simple way of detecting a trend in seasonal data. Change in averages with time is evidence of a trend in the given series, though there are more formal tests for detecting a ***trend in time series***.

4. **Other Irregular Variation (Irregular Fluctuations):**

When trend and cyclical variations are removed from a set of time series data, the residual left, which may or may not be random. Various techniques for analyzing series of this type examine to see "if irregular variation may be explained in terms of probability models such as moving average or autoregressive models, i.e. we can see if any cyclical variation is still left in the residuals. These variations occur due to sudden causes are called residual variation (irregular variation or accidental or erratic fluctuations) and are unpredictable, for example, a rise in prices of steel due to strike in the factory, accident due to failure of the break, flood, earth quick, war, etc.



(1) Long-term Trend

(2) Long-term Trend with Cyclical Variations/Movements

(3) Long-term Trend with Cyclical and Seasonal Variations/Movements

(4) Long-term Trend with Cyclical, Seasonal, and Random Variations/Movements

Models of Time Series Analysis :

In time series quantitative data are arranged in the order of their occurrence and resulting statistical series. The quantitative values are usually recorded over equal time intervals such as daily, weekly, monthly, quarterly, half-yearly, yearly, or any other measure of time.

Some examples are statistics of Industrial Production in India on a monthly basis, birth-rate figures annually, the yield on ordinary shares, and weekly wholesale price of rice, etc.

Components of Time Series

There is a different kind of forces which influence the time series analysis. Some are continuously effective while others make themselves felt at recurring time intervals. So, our first task is to divide the data and elements into components.

A time series consists of the following four components or basic elements:

1. Basic or Secular or Long-time trend;
2. Seasonal variations;
3. Business cycles or cyclical movement; and
4. Erratic or Irregular fluctuations.

These components provide a basis for the explanation of the behavior on the past time. With their help, one can predict the behavior ahead. The major tendency of each component or constituent is largely due to causal factors.

Mathematical Statements of Time Series

Some time series may not be affected by all type of variations. Some of these types of variations may affect a few time series only. Hence, while analyzing the time series, these effects are isolated. In a traditional time series analysis, we assume that any given observation consists of the trend, seasonal, cyclical and irregular movements.

Models of Time Series Analysis

The following are the two models which we generally use for the decomposition of time series into its four components. The objective is to estimate and separate the four types of variations and to bring out the relative effect of each on the overall behavior of the time series.

(1) Additive model, and

(2) Multiplicative model

1) Additive Model

In the additive model, we represent a particular observation in a time series as the sum of these four components.

i.e. $O = T + S + C + I$

where O represents the original data, T represents the trend. S represents the seasonal variations, C represents the cyclical variations and I represents the irregular variations.

In another way, we can write $Y(t) = T(t) + S(t) +C(t) + I(t)$

2) Multiplicative Model

In this model, four components have a multiplicative relationship. So, we represent a particular observation in a time series as the product of these four components:

i.e. $O = T \times S \times C \times I$

where O, T, S, C and I represents the terms as in additive model.

In another way, we can write $Y(t) = T(t) \times S(t) \times C(t) \times I(t)$

This model is the most used model in the decomposition of time series. To remove any doubt between the two models, it should be made clear that in Multiplicative model S, C, and I are indices expressed as decimal percentages whereas, in Additive model S, C and I are quantitative deviations about a trend that can be expressed as seasonal, cyclical and irregular in nature.

Example:

If in a multiplicative model.

$T = 500$, $S = 1.4$, $C = 1.20$ and $I = 0.7$

then $O=T \times S \times C \times I$

By substituting the values we get

$O = 500 \times 1.4 \times 1.20 \times 0.7 = 588$

If in additive model,

$T = 500$, $S = 100$, $C = 25$, $I = –60$

then $O = 500 + 100 + 25 – 60 = 565$

Solved Question on Models of Time Series Analysis

Q. Which model is more appropriate for time series analysis?

Solution: The assumption for the two schemes of analysis is that whereas there is no interaction among the different constituents or components under the additive scheme, such interaction is very much present in the multiplicative scheme. They do not depend on the level of the trend. With higher trends, these variations are more intensive. Though in practice the multiplicative model is the more popular, both models have their own merits. Depending on the nature of the time series analysis, they are equally acceptable.

Method of Semi Averages:

This method is very simple and relatively objective as a freehand method. In this method, we classify the time series data into two equal parts and then calculate averages for each half. If the data is for even number of years, it is easily divided into two. If the data is for an odd number of years, then the year at the middle of the time series is left and the two halves are constituted with the period on each side of the mid-year. Let us discuss the Method of Semi Averages in detail.

In this method, we can find the solution of a secular trend. For this, we have to show our time series on graph paper. For example, we can take sales on X-axis and data of production on Y-axis. Now make the original graph by plotting points on graph paper with time and value pairs. After plotting original data, we can calculate the trend line. For calculating the trend line, we will calculate semi-average.

We divide the data into two equal parts with respect to time. And then we plot the arithmetic mean of the sets of values of Y against the center of the relative time span. If the number of observations is even then the division into halves will be done easily.

But, for an odd number of observations, we will drop the middle most item, i.e. n+12 th term. We need to join these two points together through a straight line which shows the trend. The trend values can then be read from the graph corresponding to each time period.

Since extreme values greatly affect the arithmetic mean, and it is subjected to misleading values. Due to this, these trends may give distorted plots. But, if extreme values are not apparent, we may easily use and employ this method. To understand the estimation of trends, using the above mentioned two methods, consider the following working example.

Advantages:

1. This method is simple to understand as compare to other methods for measuring the secular trends.
2. Everyone who applies this method will get the same result.

Disadvantages:

1. The method assumes a straight line relationship between the plotted points without considering the fact whether that relationship exists or not.
2. If we add more data to the original data then we have to do the complete process again for the new data to get the trend values and the trend line also changes.
3.

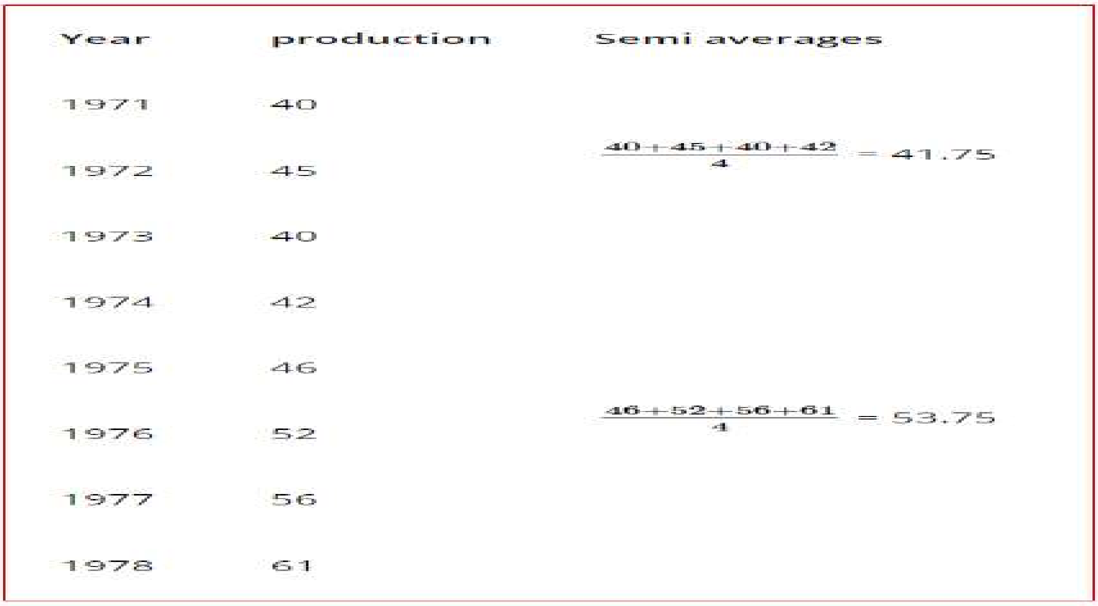
Explanation of the Method:

Here are two cases of calculating semi-average of data:

When data is even: In this case, the time series will be into two parts and then we calculate the average of each part. Suppose if we have 10 years data then we divide it into 5 -5 years and then we will calculate the first five-year average and the next five-year average after this we have to plot this on the graph paper. This will show the trend line as shown in the picture.

When data is odd: In this case, we just leave the middle data and we will follow the above-said procedure for the rest.

Solved Example on Method of Semi Averages:



Year	production	Semi averages
1971	40	
1972	45	$\frac{40+45+40+42}{4} = 41.75$
1973	40	
1974	42	
1975	46	
1976	52	$\frac{46+52+56+61}{4} = 53.75$
1977	56	
1978	61	

Thus we get two points 41.75 and 53.75 which we shall plot corresponding to their middle years i.e. 1972.5 and 1976.5. By joining these points we will obtain the required trend line.

<u>Calculation of Trend by Moving Average Method:</u>

While watching the news you might have noticed the reporter saying that the temperature of a particular city or a country has broken a record. The rainfall of some state or country has set a new bar. How can they know about it? What are the measures that they have taken and studied to say so? These are the time-series data. You all are familiar with time-series data and the various components of the time series. In this section, we will study how to calculate the trend in a set of data by the method of moving average.

A Trend in a Time Series

A time series is broadly classified into three categories of long-term fluctuations, short-term or periodic fluctuations, and random variations. A long-term variation or a trend shows the general tendency of the data to increase or decrease during a long period of time. The variation may be gradual but it is inevitably present.

Analysis of Time Series

Suppose you have a time series data. What will you do with it? How can you calculate the effect of each component for the resulting variations in it? The main problems in the analysis of time series are

- To identify the components and the net effect of whose interaction is shown by the movement of a time series, and
- To isolate, study, analyze and measure each component independently by making others constant.

Measurement of Trend by the Method of Moving Average

This method uses the concept of ironing out the fluctuations of the data by taking the means. It measures the trend by eliminating the changes or the variations by means of a moving average. The simplest of the mean used for the measurement of a trend is the arithmetic means (averages).

Moving Average

The moving average of a period (extent) *m* is a series of successive averages of *m* terms at a time. The data set used for calculating the average starts with first, second, third and etc. at a time and *m* data taken at a time.

In other words, the first average is the mean of the first *m* terms. The second average is the mean of the *m* terms starting from the second data up to $(m + 1)^{\text{th}}$ term. Similarly, the third average is the mean of the *m* terms from the third to $(m + 2)^{\text{th}}$ term and so on.

If the extent or the period, *m* is odd i.e., *m* is of the form (2k + 1), the moving average is placed against the mid-value of the time interval it covers, i.e., t = k + 1. On the other hand, if *m* is even i.e., *m* = 2k, it is placed between the two middle values of the time interval it covers, i.e., t = k and t = k + 1.

When the period of the moving average is even, then we need to synchronize the moving average with the original time period. It is done by centering the moving averages i.e., by taking the average of the two successive moving averages.

Drawbacks of Moving Average

- The main problem is to determine the extent of the moving average which completely eliminates the oscillatory fluctuations.
- This method assumes that the trend is linear but it is not always the case.
- It does not provide the trend values for all the terms.

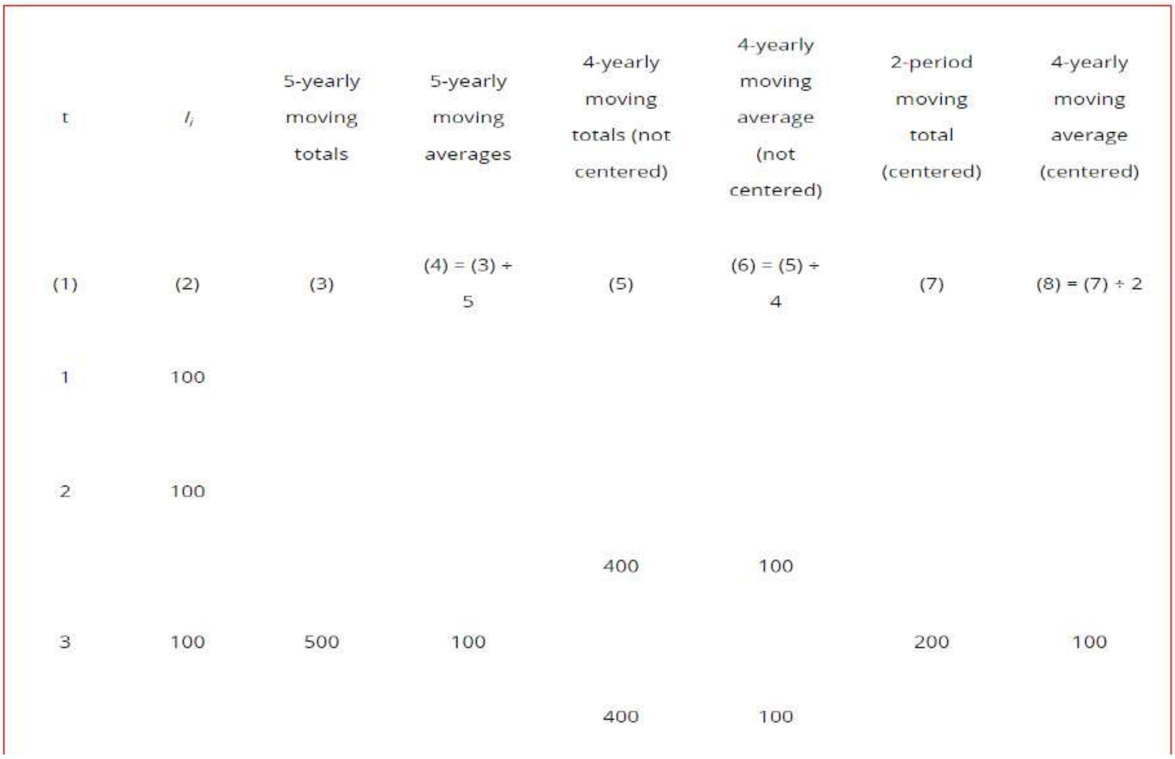
- This method cannot be used for forecasting future trend which is the main objective of the time series analysis.

Solved Example for You

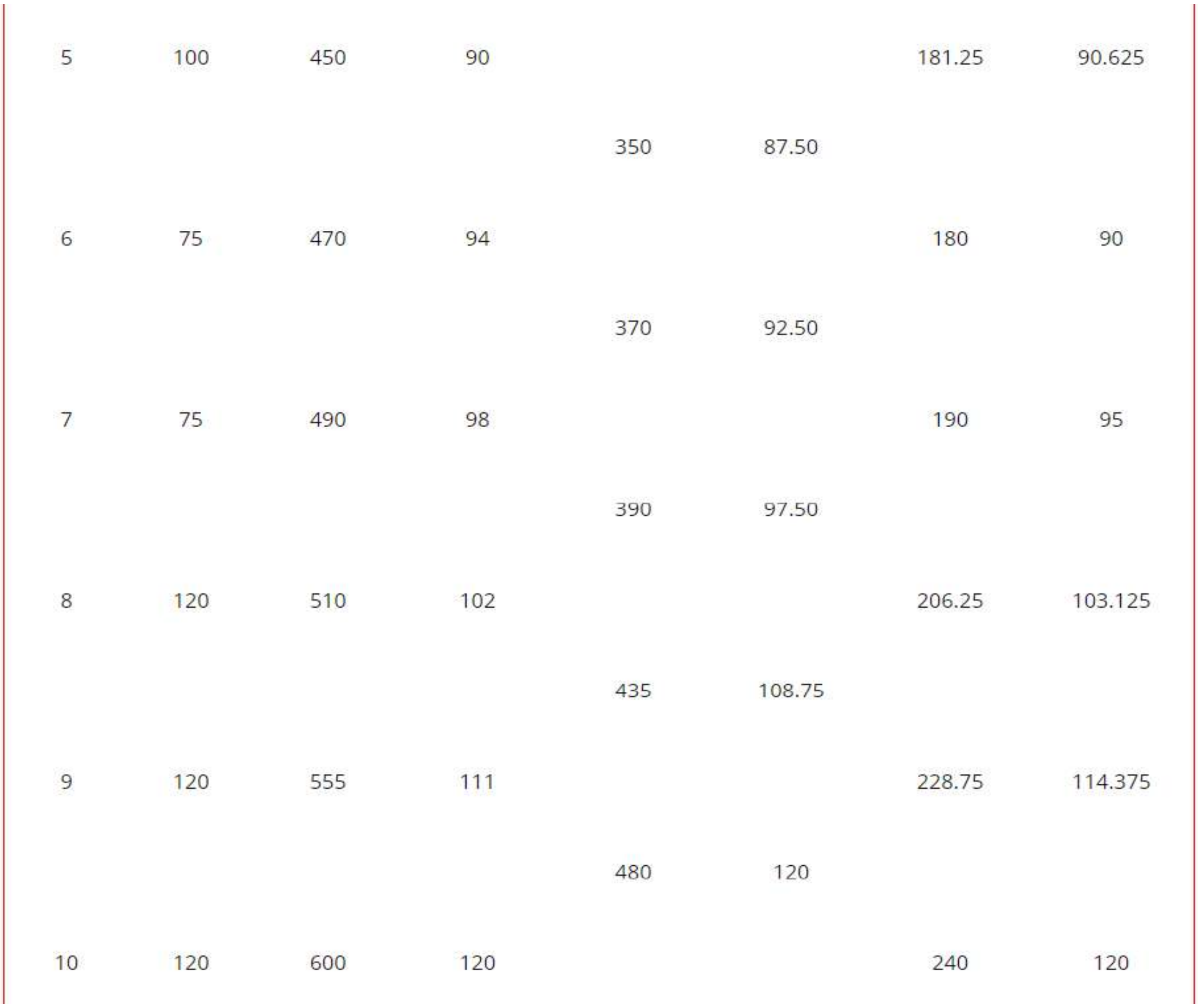
Problem: Calculate the 4-yearly and 5-yearly moving averages for the given data of the increase I_i in the population of a city for the 12 years. Make a graphic representation of it.

$$I_i = \begin{Bmatrix} 100\ ;\ t = 1, 2, \ldots, 5 \\ 75\ ;\quad t = 6, 7 \\ 120\ ; t = 8, 9, \ldots, 12 \end{Bmatrix}$$

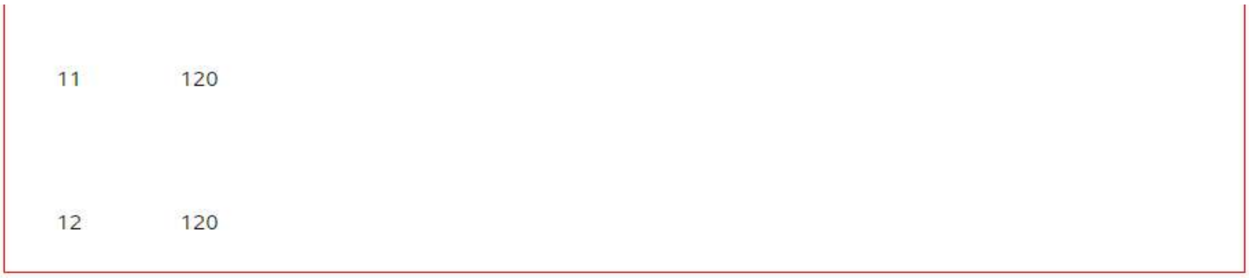
Solution:



t	I_i	5-yearly moving totals	5-yearly moving averages	4-yearly moving totals (not centered)	4-yearly moving average (not centered)	2-period moving total (centered)	4-yearly moving average (centered)
(1)	(2)	(3)	(4) = (3) ÷ 5	(5)	(6) = (5) ÷ 4	(7)	(8) = (7) ÷ 2
1	100						
2	100						
				400	100		
3	100	500	100			200	100
				400	100		

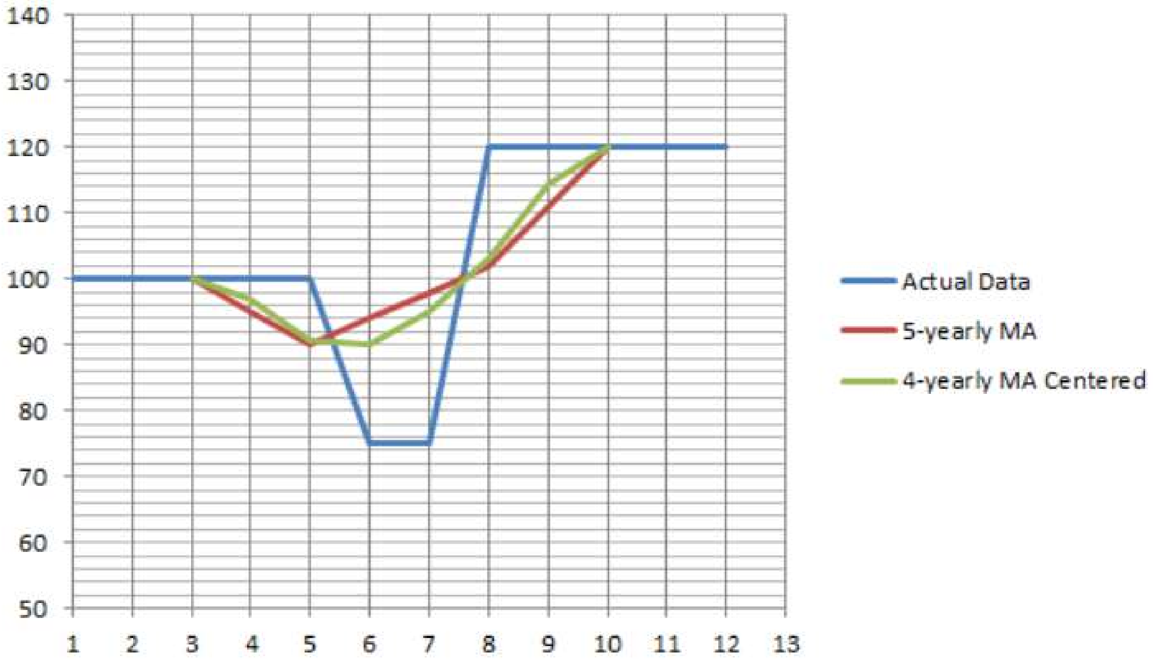


5	100	450	90			181.25	90.625
				350	87.50		
6	75	470	94			180	90
				370	92.50		
7	75	490	98			190	95
				390	97.50		
8	120	510	102			206.25	103.125
				435	108.75		
9	120	555	111			228.75	114.375
				480	120		
10	120	600	120			240	120



11	120
12	120

Here, the 4-yearly moving averages are centered so as to make the moving average coincide with the original time period. It is done by dividing the 2-period moving totals by two i.e., by taking their average. The graphic representation of the moving averages for the above data set is



Method of Least Squares:

During Time Series analysis we come across with variables, many of them are dependent upon others. It is often required to find a relationship between two or more variables. Least Square is the method for finding the best fit of a set of data points. It minimizes the sum of the residuals of points from the plotted curve. It gives the trend line of best fit to a time series data. This method is most widely used in time series analysis. Let us discuss the Method of Least Squares in detail.

Method of Least Squares

Each point on the fitted curve represents the relationship between a known independent variable and an unknown dependent variable.

In general, the least squares method uses a straight line in order to fit through the given points which are known as the method of linear or ordinary least squares. This line is termed as the line of best fit from which the sum of squares of the distances from the points is minimized.

Equations with certain parameters usually represent the results in this method. The method of least squares actually defines the solution for the minimization of the sum of squares of deviations or the errors in the result of each equation.

The least squares method is used mostly for data fitting. The best fit result minimizes the sum of squared errors or residuals which are said to be the differences between the observed or experimental value and corresponding fitted value given in the model. There are two basic kinds of the least squares methods – ordinary or linear least squares and nonlinear least squares.

Mathematical Representation

It is a mathematical method and with it gives a fitted trend line for the set of data in such a manner that the following two conditions are satisfied.

1. The sum of the deviations of the actual values of Y and the computed values of Y is zero.
2. The sum of the squares of the deviations of the actual values and the computed values is least.

This method gives the line which is the line of best fit. This method is applicable to give results either to fit a straight line trend or a parabolic trend.

The method of least squares as studied in time series analysis is used to find the trend line of best fit to a time series data.

Secular Trend Line

The secular trend line (Y) is defined by the following equation:

$Y = a + b X$

Where, Y = predicted value of the dependent variable

a = Y-axis intercept i.e. the height of the line above origin (when X = 0, Y = a)

b = slope of the line (the rate of change in Y for a given change in X)

When b is positive the slope is upwards, when b is negative, the slope is downwards

X = independent variable (in this case it is time)

To estimate the constants a and b, the following two equations have to be solved simultaneously:

$\Sigma Y = na + b\,\Sigma X$
$\Sigma XY = a\Sigma X + b\Sigma X^2$

To simplify the calculations, if the midpoint of the time series is taken as origin, then the negative values in the first half of the series balance out the positive values in the second half so that ΣX = 0. In this case, the above two normal equations will be as follows:

$\Sigma Y = na$

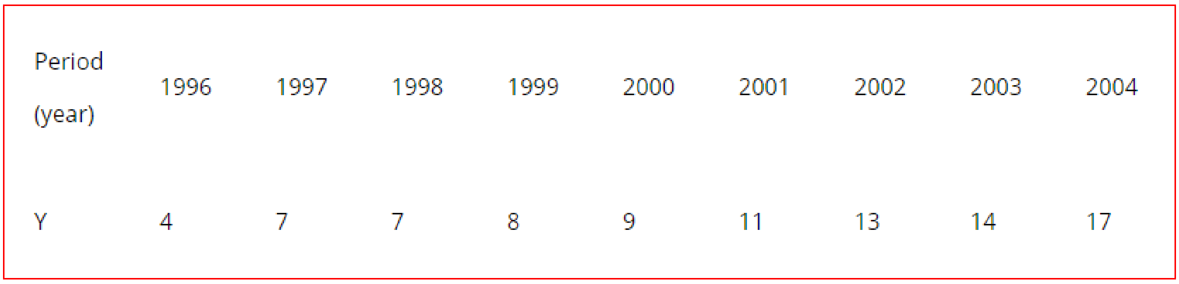
$\Sigma XY = b\Sigma X^2$

In such a case the values of a and b can be calculated as under:

Since $\Sigma Y = na$
$a = \sum Yn$
Since, $\Sigma XY = b\Sigma X^2$

Solved Example on Method of Least Squares:

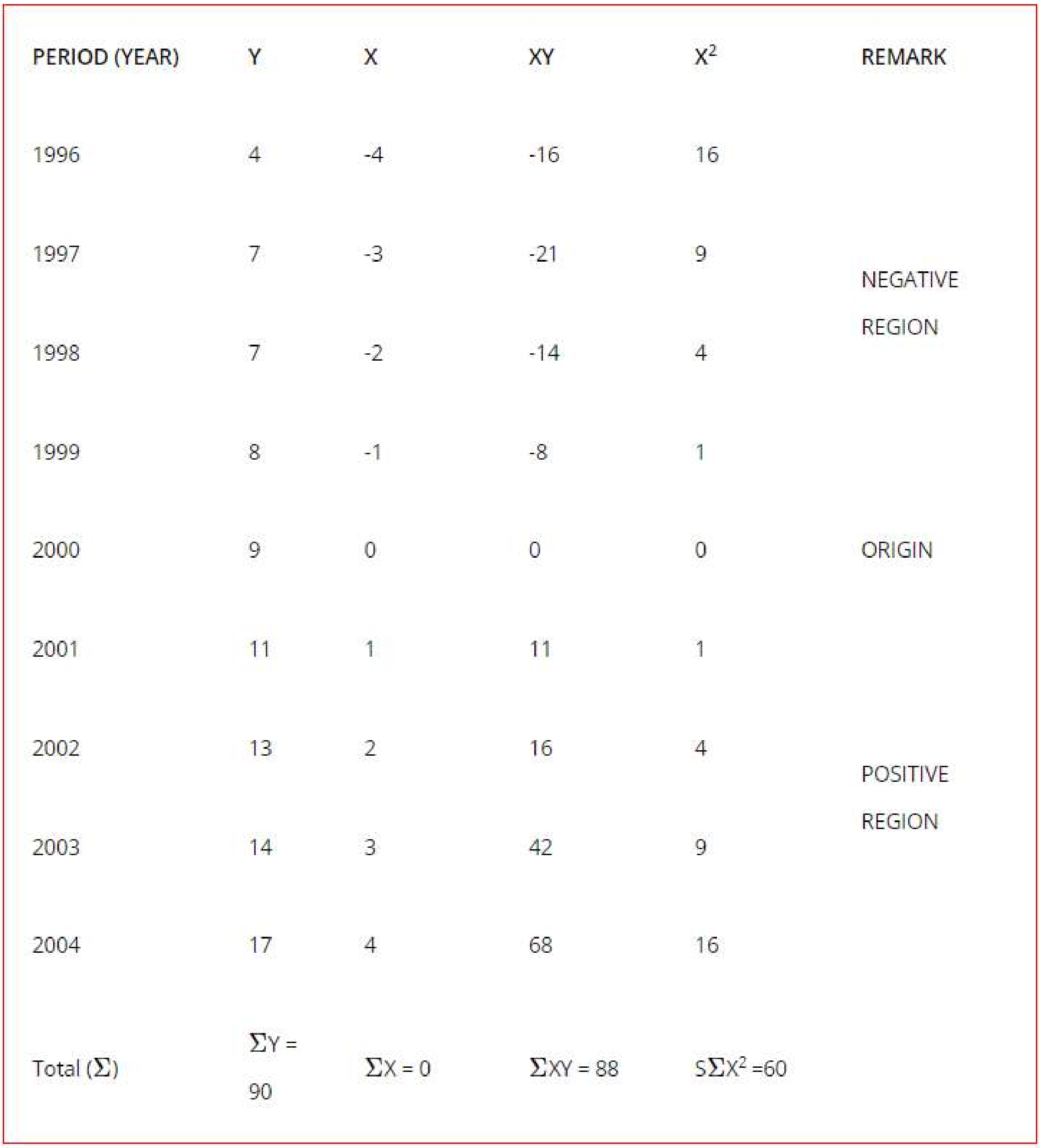
Fit a straight line trend on the following data using the Least Squares Method



Period (year)	1996	1997	1998	1999	2000	2001	2002	2003	2004
Y	4	7	7	8	9	11	13	14	17

Solution:

Total of 9 observations are there. So, the origin is taken at the Year 2000 for which X is assumed to be 0.



PERIOD (YEAR)	Y	X	XY	X^2	REMARK
1996	4	-4	-16	16	
1997	7	-3	-21	9	NEGATIVE REGION
1998	7	-2	-14	4	
1999	8	-1	-8	1	
2000	9	0	0	0	ORIGIN
2001	11	1	11	1	
2002	13	2	16	4	POSITIVE REGION
2003	14	3	42	9	
2004	17	4	68	16	
Total (Σ)	ΣY = 90	ΣX = 0	ΣXY = 88	SΣX^2 =60	

From the table we find that value of n is 9, value of ΣY is 90, value of ΣX is 0, value of ΣXY is 88 and value of ΣX^2 is 60 .

Substituting these values in the two given equations,

a = 909 or a = 10
b = 8860 or b = 1.47
Trend equation is : Y = 10 + 1.47 X

Methods of constructing seasonal indices:

Seasonal variation:

Seasonal variations are fluctuations within a year over different seasons.

Estimation of seasonal variations requires that the time series data are recorded at even intervals such as quarterly, monthly, weekly or daily, depending on the nature of the time series. Changes due to seasons, weather conditions and social customs are the primary causes of seasonal variations. The main objective of the measurement of seasonal variation is to study their effect and isolate them from the trend.

Methods of constructing seasonal indices

There are four methods of constructing seasonal indices.

1. Simple averages method

2. Ratio to trend method

3. Percentage moving average method

4. Link relatives method

Among these, we shall discuss the construction of seasonal index by the first method only.

Simple Averages Method

Under this method, the time series data for each of the 4 seasons (for quarterly data) of a particular year are expressed as percentages to the seasonal average for that year.

The percentages for different seasons are averaged over the years by using simple average.

The resulting percentages for each of the 4 seasons then constitute the required seasonal indices.

Method of calculating seasonal indices:

(i) The data is arranged season-wise

(ii) The data for all the 4 seasons are added first for all the years and the seasonal averages for each year is computed.

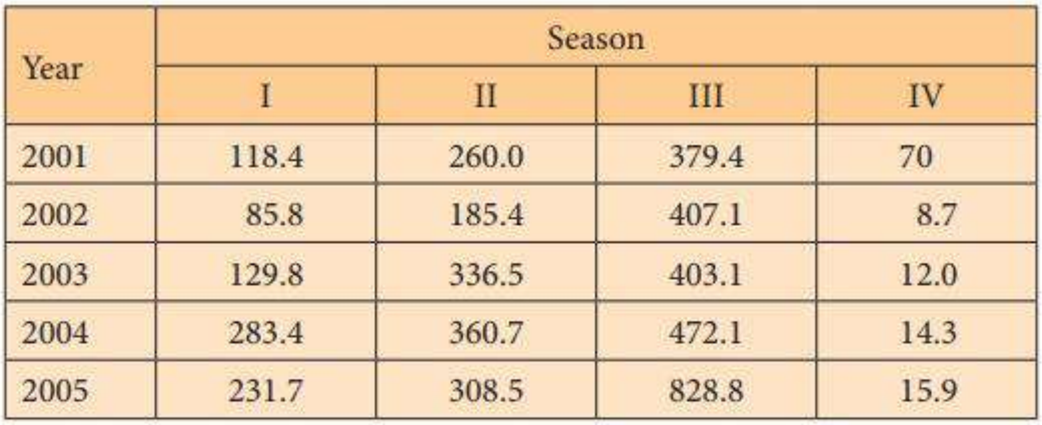
(iii) The average of seasonal averages is calculated

(*i.e.*, Grand average = Total of seasonal averages /number of years).

(iv) The seasonal average for each year is divided by the corresponding grand average and the results are expressed in percentages and these are called seasonal indices.

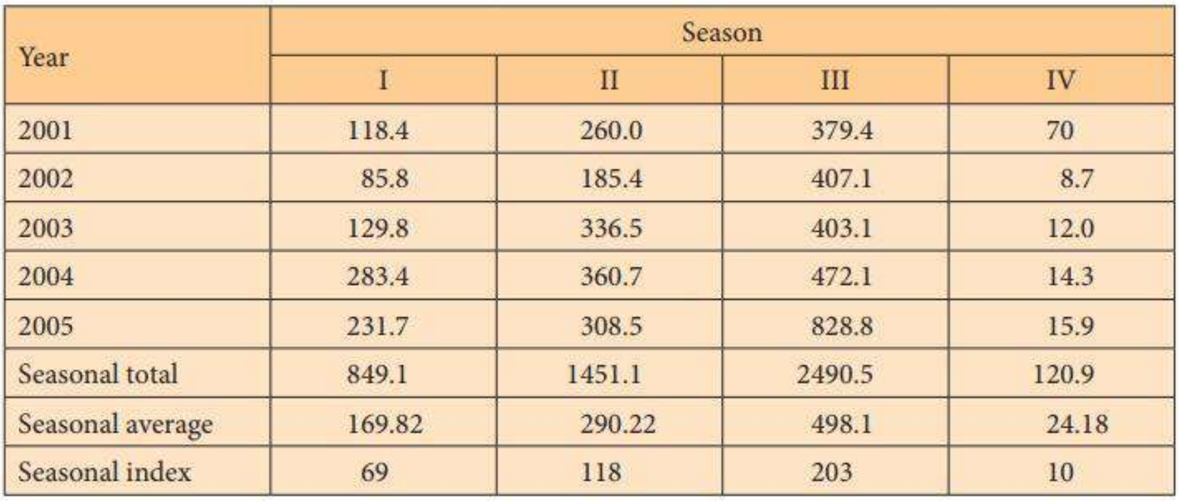
Example 7.9

Calculate the seasonal indices for the rain fall (in mm) data in Tamil Nadu given below by simple average method



Year	Season			
	I	II	III	IV
2001	118.4	260.0	379.4	70
2002	85.8	185.4	407.1	8.7
2003	129.8	336.5	403.1	12.0
2004	283.4	360.7	472.1	14.3
2005	231.7	308.5	828.8	15.9

Solution:



Year	Season			
	I	II	III	IV
2001	118.4	260.0	379.4	70
2002	85.8	185.4	407.1	8.7
2003	129.8	336.5	403.1	12.0
2004	283.4	360.7	472.1	14.3
2005	231.7	308.5	828.8	15.9
Seasonal total	849.1	1451.1	2490.5	120.9
Seasonal average	169.82	290.22	498.1	24.18
Seasonal index	69	118	203	10

$$\text{Grand Average} = \frac{\text{Total of seasonal averages}}{4}$$

$$= \frac{169.82 + 290.22 + 498.1 + 24.18}{4}$$

$$= \frac{982.32}{4} = 245.58$$

$$\text{Seasonal Index} = \frac{\text{Seasonal average}}{\text{Grand average}} \times 100$$

$$\text{Seasonal Index for Season I} = \frac{169.82}{245.58} \times 100 = 69.15 \approx 69$$

$$\text{Seasonal Index for Season II} = \frac{290.22}{245.58} \times 100 = 118.18 \approx 118$$

$$\text{Seasonal Index for Season III} = \frac{498.1}{245.58} \times 100 = 202.83 \approx 203$$

$$\text{Seasonal Index for Season IV} = \frac{24.18}{245.58} \times 100 = 9.85 \approx 10$$

RATIO TO MOVING AVERAGE METHOD:

The ratio to moving average is the most commonly used method of measuring seasonal variations. This method assumes the presence of all the four components of a time series. Various steps in the computation of seasonal indices are as follows:

1. Compute the moving averages with period equal to the period of seasonal variations. This would eliminate the seasonal components and minimize the effect of random component. The resulting moving averages would consist of trend, cyclical and random components.

2. The original values, for each quarter (or month) are divided by the respective moving average figures and the ratio is expressed as a percentage, i.e. SR” = Y / M. A = TCSR / TCR’, where R’ and R” denote the changed random components.
3. Finally, the random component R” is eliminated by the method of simple averages.

MERITS AND DEMERITS This method assumes that all the four components of a time series are present and, therefore, widely used for measuring seasonal variations. However, the seasonal variations are not completely eliminated if the cycles of these variations are not of regular nature. Further, some information is always lost at ends of the time series.

RATIO TO TREND METHOD:

This method is used when then cyclical variations are absent from the data, i.e. the time series variable Y consists of trend, seasonal and random components. Using symbols, we can write Y = T. S .R Various steps in the computation of seasonal indices are:

1. Obtain the trend values for each month or quarter, etc, by the method of least squares.

2. Divide the original values by the corresponding trend values. This would eliminate trend values from the data.

3. To get figures in percentages, the quotients are multiplied by 100. Thus, we have three equations:

$$Y/\,T \times 100$$
$$T.\,S.\,R\,/\,T \times 100$$
$$S.\,R \times 100$$

MERITS AND DEMERITS:

It is an objective method of measuring seasonal variations. However, it is very complicated and doesn't work if cyclical variations are present.

4. **LINK RELATIVES METHOD :**

This method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern. The link relatives are percentages of the current period (quarter or month) as compared with the previous period. With the computations of the link relatives and their average, the effect of cyclical and the random components is minimized. Further, the trend gets eliminated in the process of adjustment of chain relatives.

The following steps are involved in the computation of seasonal indices by this method: 1. Compute the Link Relative (L.R.) of each period by dividing the figure of that period with the figure of previous period. For example, Link relative of 3rd quarter=figure of 3rd quarter / figure of 2nd quarter ×100.

2. Obtain the average of link relatives of a given quarter (or month) of various years. A.M. or Md can be used for this purpose. Theoretically, the later is preferable because the former gives undue importance to extreme items.

3. These averages are converted into chained relatives by assuming the chained relative of the first quarter (or month) equal to 100. The chained relative (C.R.) for the current period (quarter or month) = C.R. of the previous period ×L.R. of the current period / 100.

4.Compute the C.R. of the first quarter (or month) on the basis of the last quarter (or month). This is given by C.R. of the last quarter (month) × average L.R. of the first quarter (month) / 100

a. This value, in general, is different from 100 due to long term trend in the data. The chained relatives, obtained above, are to be adjusted for the effect of this trend.

The adjustment factor

d=14new C.R for 1st quater-100for quaterly data d=112

new C.R for 1st month-100for monthly data

b. On the assumption that the trend is linear d, 2d, 3d, etc, is respectively subtracted from the 2nd , 3 rd , 4th , etc quarter (or month).

5. Express the adjusted chained relatives as a percentage of their average to obtain seasonal indices.

6. Make sure that the sum of these indices is 400 for quarterly data and 1200 for monthly data.

MERITS AND DEMERITS:

This method is less complicated than the ratio to moving average and the ratio to trend methods. However, this method is based upon the assumption of a linear trend which may not always hold true.

Chapter – 5

INDEX NUMBERS

Meaning of Index Numbers:

Index number is a technique of measuring changes in a variable or group of variables with respect to time, geographical location or other characteristics. There can be various types of index numbers.

Features of Index Numbers:

The following are the main features of index numbers:

(i) Index numbers are a special type of average. Whereas mean, median and mode measure the absolute changes and are used to compare only those series which are expressed in the same units, the technique of index numbers is used to measure the relative changes in the level of a phenomenon where the measurement of absolute change is not possible and the series are expressed in different types of items.

(ii) Index numbers are meant to study the changes in the effects of such factors which cannot be measured directly. For example, the general price level is an imaginary concept and is not capable of direct measurement. But, through the technique of index numbers, it is possible to have an idea of relative changes in the general level of prices by measuring relative changes in the price level of different commodities.

(iii) The technique of index numbers measures changes in one variable or group of related variables. For example, one variable can be the price of wheat, and group of variables can be the price of sugar, the price of milk and the price of rice.

(iv) The technique of index numbers is used to compare the levels of a phenomenon on a certain date with its level on some previous date (e.g., the price level in 1980 as compared to that in 1960 taken as the base year) or the levels of a phenomenon at different places on the same date (e.g., the price level in India in 1980 in comparison with that in other countries in 1980).

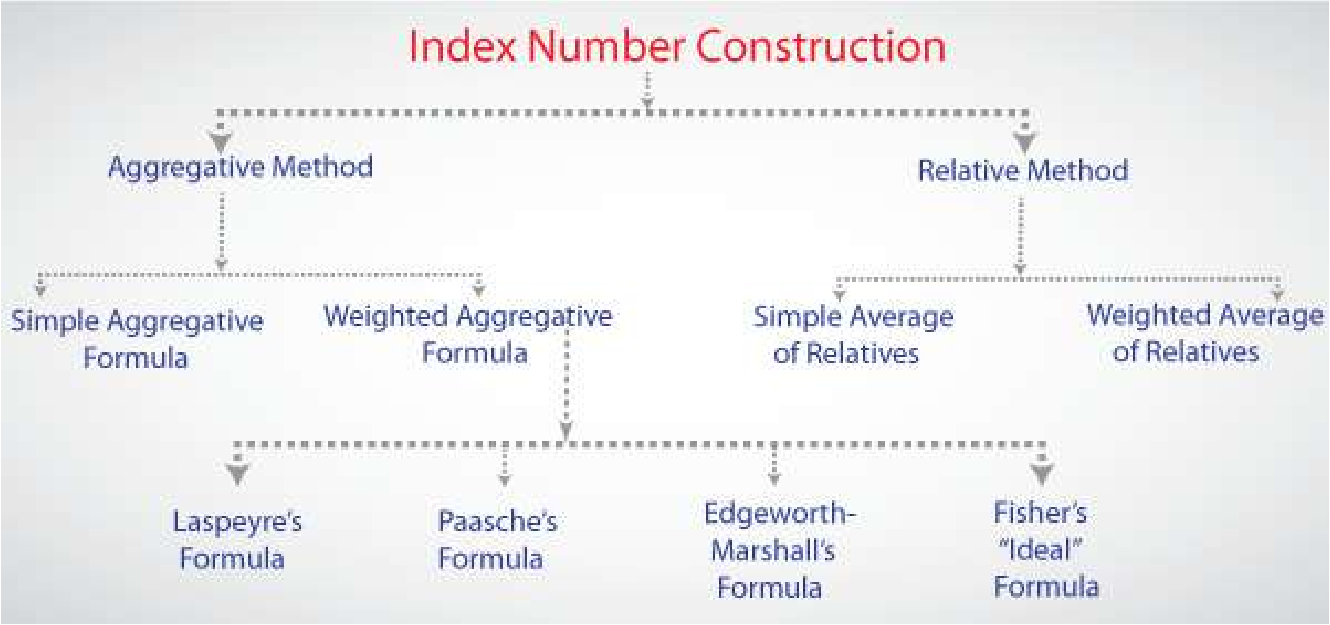
Types of Index Numbers:

Simple Index Number: A simple index number is a number that measures a relative change in a single variable with respect to a base. These type of Index numbers are constructed from a single item only.

Composite Index Number: A composite index number is a number that measures an average relative changes in a group of relative variables with respect to a base. A composite index number is built from changes in a number of different items.

Price index Numbers: Price index numbers measure the relative changes in prices of a commodity between two periods. Prices can be either retail or wholesale. Price index number are useful to comprehend and interpret varying economic and business conditions over time.

Quantity Index Numbers: These types of index numbers are considered to measure changes in the physical quantity of goods produced, consumed or sold of an item or a group of items. Methods of constructing index numbers: There are two methods to construct index numbers: Price relative and aggregate methods.


Index Number Construction
Aggregative Method
Relative Method
Simple Aggregative Formula
Weighted Aggregative Formula
Simple Average of Relatives
Weighted Average of Relatives
Laspeyre's Formula
Paasche's Formula
Edgeworth-Marshall's Formula
Fisher's "Ideal" Formula

In aggregate methods, the aggregate price of all items in a given year is expressed as a percentage of same in the base year, giving the index number.

Difficulties faced in the Construction of Index Numbers

There are many difficulties faced in the construction of index numbers. They are discussed as under:

1. *Difficulties in the Selection of the Base Period:*

The first difficulty relates to the selection of the base year. The base year should be normal. But it is difficult to determine a truly normal year. Moreover, what may be the normal year today may become an abnormal year after some period. Therefore, it is not advisable to have the same year as the base period for a number of years. Rather, it should be changed after ten years or so. But there is no fixed rule for this.

2. Difficulties in the Selection of Commodities:

The selection of representative commodities for the index number is another difficulty. The choice of representative commodities is not an easy matter. They have to be selected from a wide range of commodities which the majority of people consume. Again, what were representative commodities some ten years ago may not be representative today. The consumption pattern of consumers might change and thereby make the index number useless. So the choice of representative commodities presents real difficulties.

3. Difficulties in the Collection of Prices:

Another difficulty is that of collecting adequate and accurate prices. It is often not possible to get them from the same source or place. Further, the problem of choice between wholesale and retail prices arises. There are much variations in the retail prices. Therefore, index numbers are based on wholesale prices.

4. Arbitrary Assigning of Weights:

In calculating weighted price index, a number of difficulties arise. The problem is to give different weights to commodities. The selection of higher weight for one commodity and a lower weight for another is simply arbitrary. There is no set rule and it entirely depends on the investigator. Moreover, the same commodity may have different importance for different consumers. The importance of commodities also changes with the change in the tastes and incomes of consumers and also with the passage of time. Therefore, weights are to be revised from time to time and not fixed arbitrarily.

5. Difficulty of Selecting the Method of Averaging:

Another difficulty is to select an appropriate method of calculating averages. There are a number of methods which can be used for this purpose. But all methods give different results from one another. It is, therefore, difficult to decide which method to choose.

6.Difficulties Arising from Changes Overtime:

In the present times, changes in the nature of commodities are taking place continuously overtime due to technological changes. As a result, new commodities are introduced and people start consuming them in place of the old ones. Moreover, prices of commodities might also change with technical changes. They may fall. But new commodities are not entered into the list of commodities in preparing the index numbers. This makes the index numbers based on old commodities unreal.

7. Not All Purpose:

An index number constructed for a particular purpose cannot be used for some other purpose. For instance, a cost of living index number for industrial workers cannot be used to measure the cost of living of agricultural workers. Thus there are no all purpose index numbers.

8. International Comparisons not Possible:

International price comparisons are not possible with index numbers. The commodities consumed and included in the construction of an index number differ from country to country. For instance, meat, eggs, cars, and electrical appliance are included in the price index of advanced countries whereas they are not included in that of backward countries. Similarly, weights assigned to commodities are also different. Thus international comparisons of index numbers are not possible.

9. Comparisons of Different Places not Possible:

Even if different places within a country are taken, it is not possible to apply the same index number to them. This is because of differences in the consumption habits of people. People living in the northern region consume different commodities than those consumed by the people in the south of India. It is, therefore, not right to apply the same index number to both.

10. Not Applicable to an Individual:

An index number is not applicable to an individual belonging to a group for which it is constructed. If an index number shows a rise in the price level, an individual may not be affected by it. This is because an index number reflects averages.

Conclusion:

It may be concluded from the difficulties and limitations of index numbers that index numbers are at best approximations to measure changes in the value of money. However, these difficulties become less significant if index numbers are constructed for short intervals. This is because habits, tastes, techniques of production and the qualities of commodities entering into a price index number do not change during the short period.

METHODS OF CALCULATING INDEX NUMBERS:

There are two methods of computing the index numbers:

(a) Simple index number and
(b) Weighted index number.

Simple index number again can be constructed either by – (i) Simple aggregate method, or by (ii) simple average of price relative’s method.

Similarly, weighted index number can be constructed either by (i) weighted aggregative method, or by (ii) weighted average of price relative’s method.

The choice of method depends upon the availability of data, degree of accuracy required and the purpose of the study.

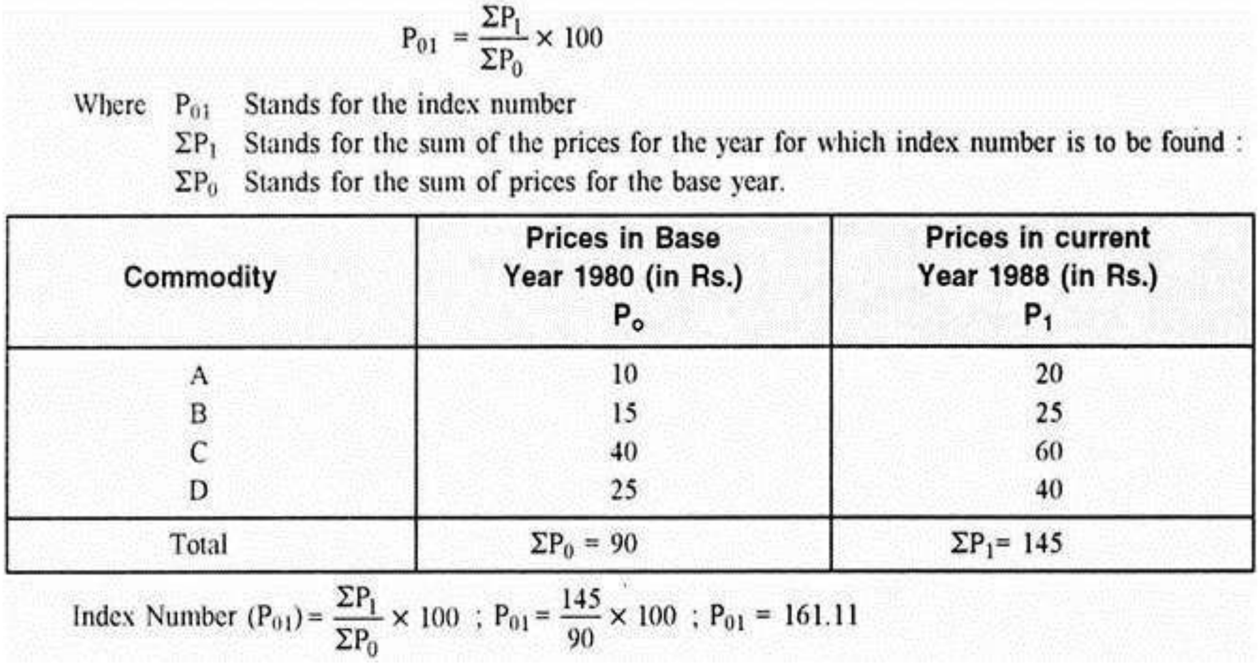
Construction of Price Index Numbers (Formula and Examples):

Construction of price index numbers through various methods can be understood with the help of the following examples:

1. Simple Aggregative Method:

In this method, the index number is equal to the sum of prices for the year for which index number is to be found divided by the sum of actual prices for the base year.

The formula for finding the index number through this method is as follows:



$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

Where P_{01} Stands for the index number

ΣP_1 Stands for the sum of the prices for the year for which index number is to be found :

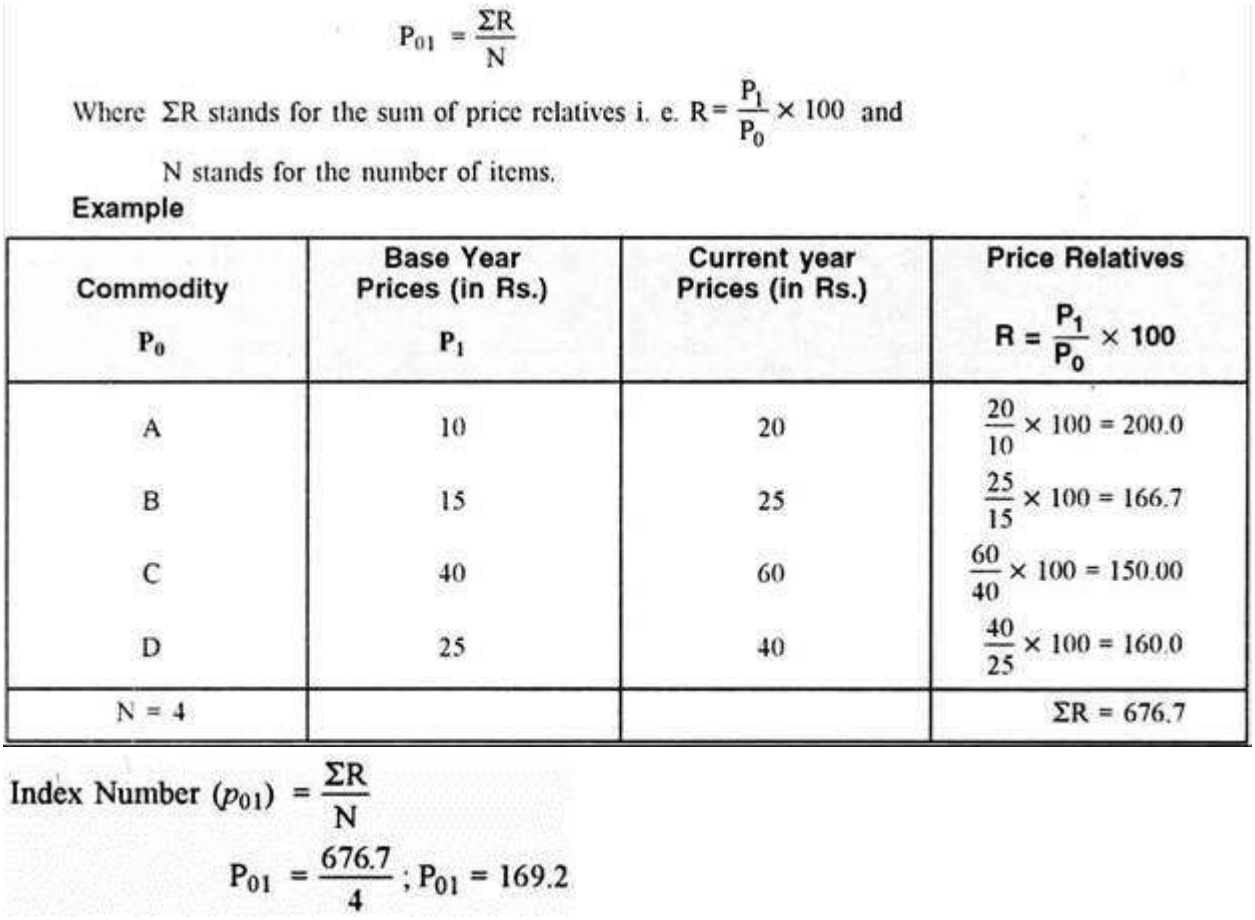
ΣP_0 Stands for the sum of prices for the base year.

Commodity	Prices in Base Year 1980 (in Rs.) P_o	Prices in current Year 1988 (in Rs.) P_1
A	10	20
B	15	25
C	40	60
D	25	40
Total	$\Sigma P_0 = 90$	$\Sigma P_1 = 145$

Index Number $(P_{01}) = \frac{\Sigma P_1}{\Sigma P_0} \times 100$; $P_{01} = \frac{145}{90} \times 100$; $P_{01} = 161.11$

2. Simple Average of Price Relatives Method:

In this method, the index number is equal to the sum of price relatives divided by the number of items and is calculated by using the following formula:



$$P_{01} = \frac{\Sigma R}{N}$$

Where ΣR stands for the sum of price relatives i. e. $R = \frac{P_1}{P_0} \times 100$ and

N stands for the number of items.

Example

Commodity P_0	Base Year Prices (in Rs.) P_1	Current year Prices (in Rs.)	Price Relatives $R = \frac{P_1}{P_0} \times 100$
A	10	20	$\frac{20}{10} \times 100 = 200.0$
B	15	25	$\frac{25}{15} \times 100 = 166.7$
C	40	60	$\frac{60}{40} \times 100 = 150.00$
D	25	40	$\frac{40}{25} \times 100 = 160.0$
N = 4			$\Sigma R = 676.7$

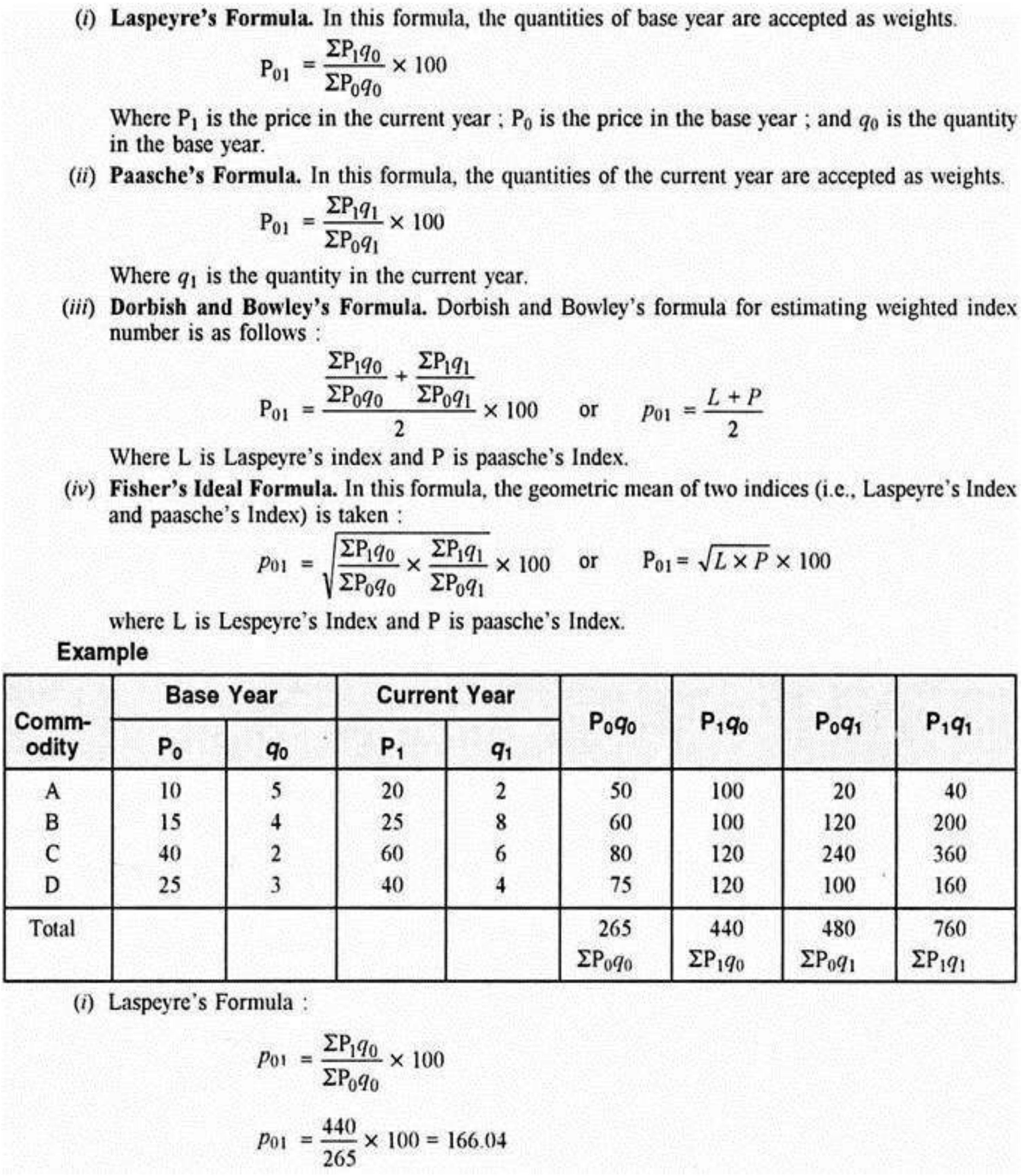
$$\text{Index Number } (p_{01}) = \frac{\Sigma R}{N}$$

$$P_{01} = \frac{676.7}{4}\ ; P_{01} = 169.2$$

3. Weighted Aggregative Method:

In this method, different weights are assigned to the items according to their relative importance. Weights used are the quantity weights. Many formulae have been developed to estimate index numbers on the basis of quantity weights.

Some of them are explained below:



(*i*) **Laspeyre's Formula.** In this formula, the quantities of base year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

Where P_1 is the price in the current year ; P_0 is the price in the base year ; and q_0 is the quantity in the base year.

(*ii*) **Paasche's Formula.** In this formula, the quantities of the current year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

Where q_1 is the quantity in the current year.

(*iii*) **Dorbish and Bowley's Formula.** Dorbish and Bowley's formula for estimating weighted index number is as follows :

$$P_{01} = \frac{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 \quad \text{or} \quad p_{01} = \frac{L + P}{2}$$

Where L is Laspeyre's index and P is paasche's Index.

(*iv*) **Fisher's Ideal Formula.** In this formula, the geometric mean of two indices (i.e., Laspeyre's Index and paasche's Index) is taken :

$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100 \quad \text{or} \quad P_{01} = \sqrt{L \times P} \times 100$$

where L is Lespeyre's Index and P is paasche's Index.

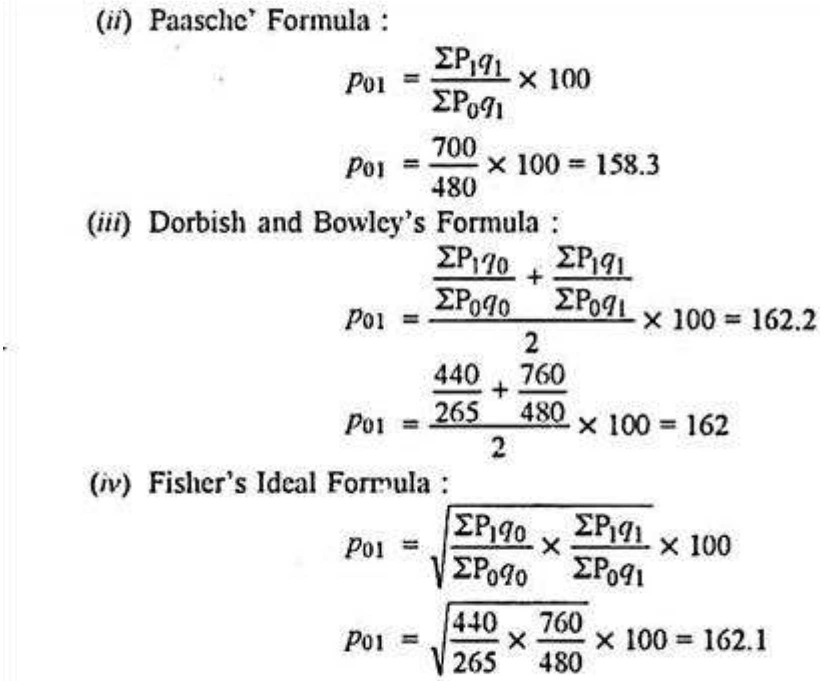
Example

Comm-odity	Base Year		Current Year		P_0q_0	P_1q_0	P_0q_1	P_1q_1
	P_0	q_0	P_1	q_1				
A	10	5	20	2	50	100	20	40
B	15	4	25	8	60	100	120	200
C	40	2	60	6	80	120	240	360
D	25	3	40	4	75	120	100	160
Total					265 ΣP_0q_0	440 ΣP_1q_0	480 ΣP_0q_1	760 ΣP_1q_1

(*i*) Laspeyre's Formula :

$$p_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

$$p_{01} = \frac{440}{265} \times 100 = 166.04$$



(*ii*) Paasche' Formula :

$$p_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

$$p_{01} = \frac{700}{480} \times 100 = 158.3$$

(*iii*) Dorbish and Bowley's Formula :

$$p_{01} = \frac{\dfrac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \dfrac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 = 162.2$$

$$p_{01} = \frac{\dfrac{440}{265} + \dfrac{760}{480}}{2} \times 100 = 162$$

(*iv*) Fisher's Ideal Formula :

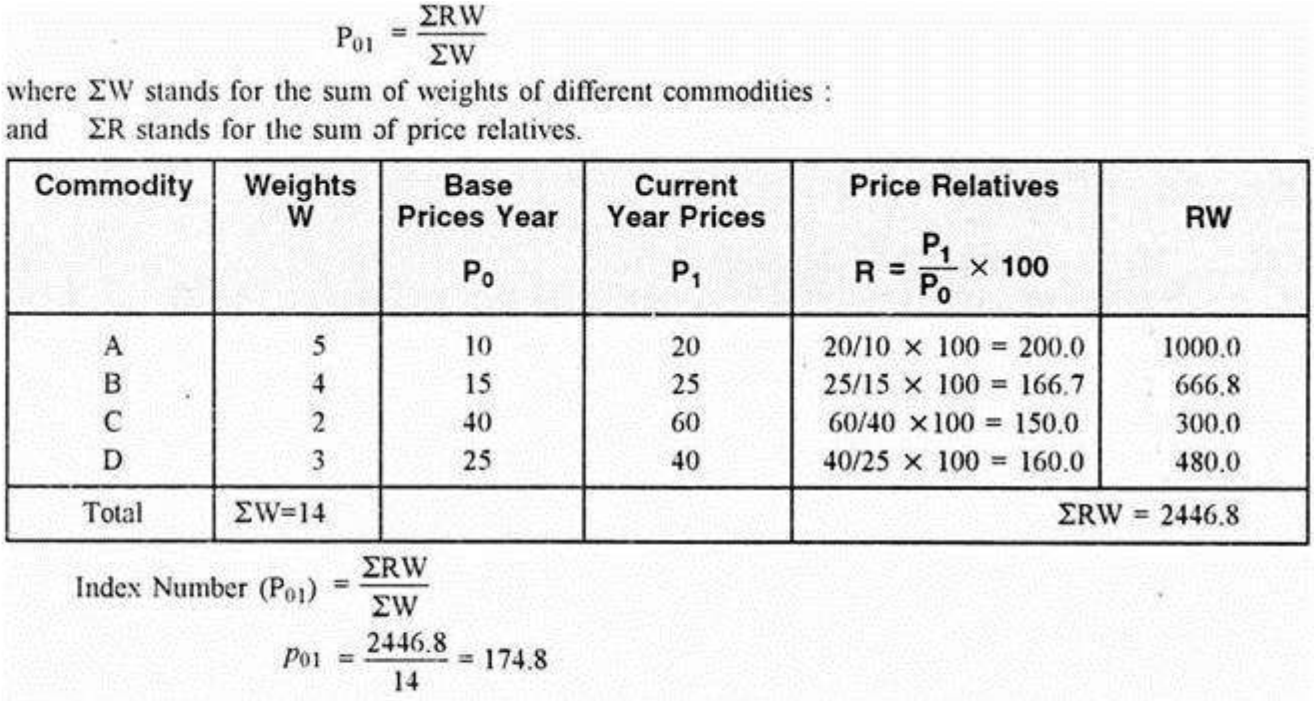
$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100$$

$$p_{01} = \sqrt{\frac{440}{265} \times \frac{760}{480}} \times 100 = 162.1$$

4. Weighted Average of Relatives Method:

In this method also different weights are used for the items according to their relative importance.

The price index number is found out with the help of the following formula:



$$P_{01} = \frac{\Sigma RW}{\Sigma W}$$

where ΣW stands for the sum of weights of different commodities :

and ΣR stands for the sum of price relatives.

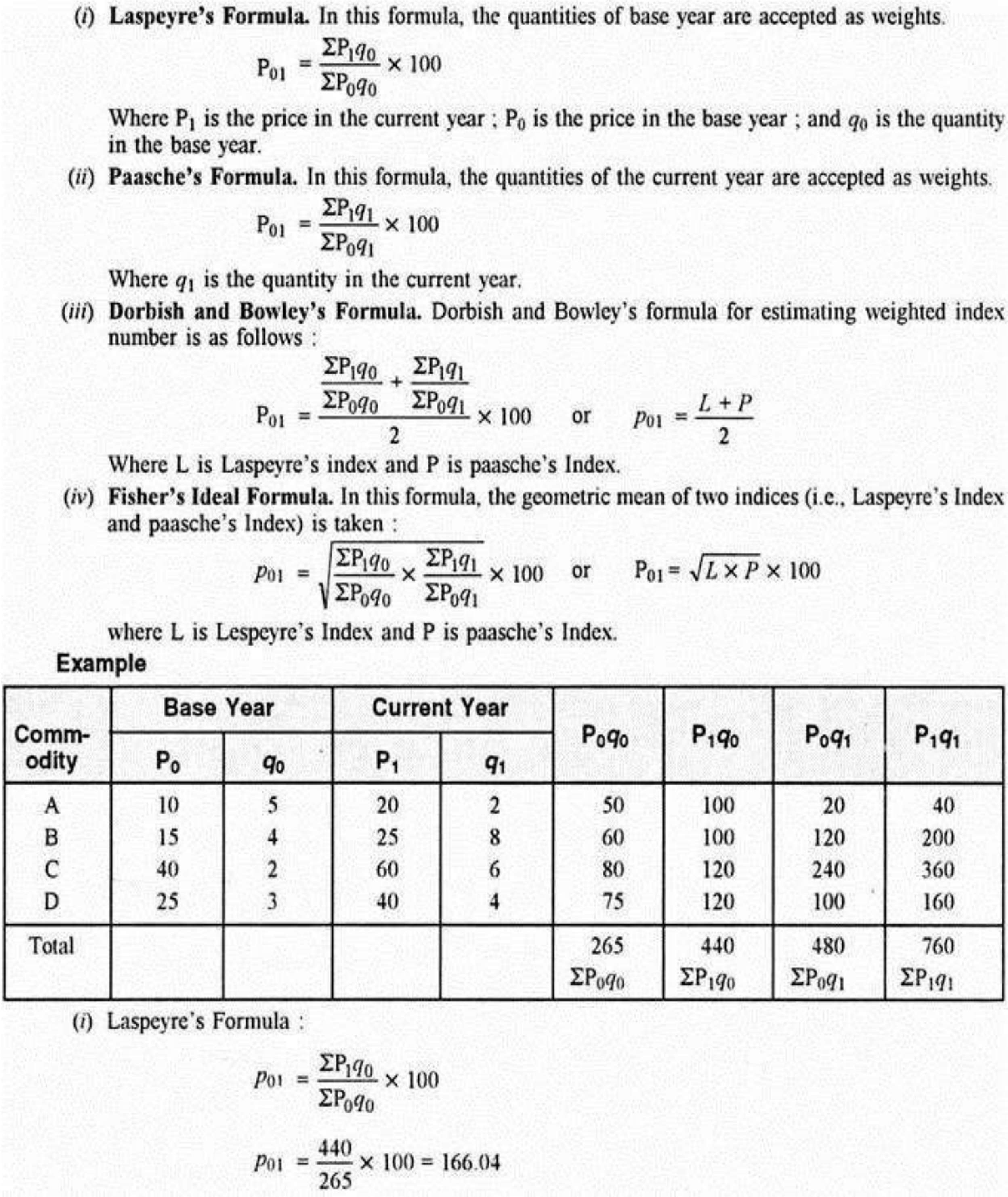
Commodity	Weights W	Base Prices Year P_0	Current Year Prices P_1	Price Relatives $R = \frac{P_1}{P_0} \times 100$	RW
A	5	10	20	20/10 × 100 = 200.0	1000.0
B	4	15	25	25/15 × 100 = 166.7	666.8
C	2	40	60	60/40 ×100 = 150.0	300.0
D	3	25	40	40/25 × 100 = 160.0	480.0
Total	ΣW=14				ΣRW = 2446.8

$$\text{Index Number } (P_{01}) = \frac{\Sigma RW}{\Sigma W}$$

$$p_{01} = \frac{2446.8}{14} = 174.8$$

In this method, different weights are assigned to the items according to their relative importance. Weights used are the quantity weights. Many formulae have been developed to estimate index numbers on the basis of quantity weights.

Some of them are explained below:



(*i*) **Laspeyre's Formula.** In this formula, the quantities of base year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

Where P_1 is the price in the current year ; P_0 is the price in the base year ; and q_0 is the quantity in the base year.

(*ii*) **Paasche's Formula.** In this formula, the quantities of the current year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

Where q_1 is the quantity in the current year.

(*iii*) **Dorbish and Bowley's Formula.** Dorbish and Bowley's formula for estimating weighted index number is as follows :

$$P_{01} = \frac{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 \quad \text{or} \quad p_{01} = \frac{L + P}{2}$$

Where L is Laspeyre's index and P is paasche's Index.

(*iv*) **Fisher's Ideal Formula.** In this formula, the geometric mean of two indices (i.e., Laspeyre's Index and paasche's Index) is taken :

$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100 \quad \text{or} \quad P_{01} = \sqrt{L \times P} \times 100$$

where L is Lespeyre's Index and P is paasche's Index.

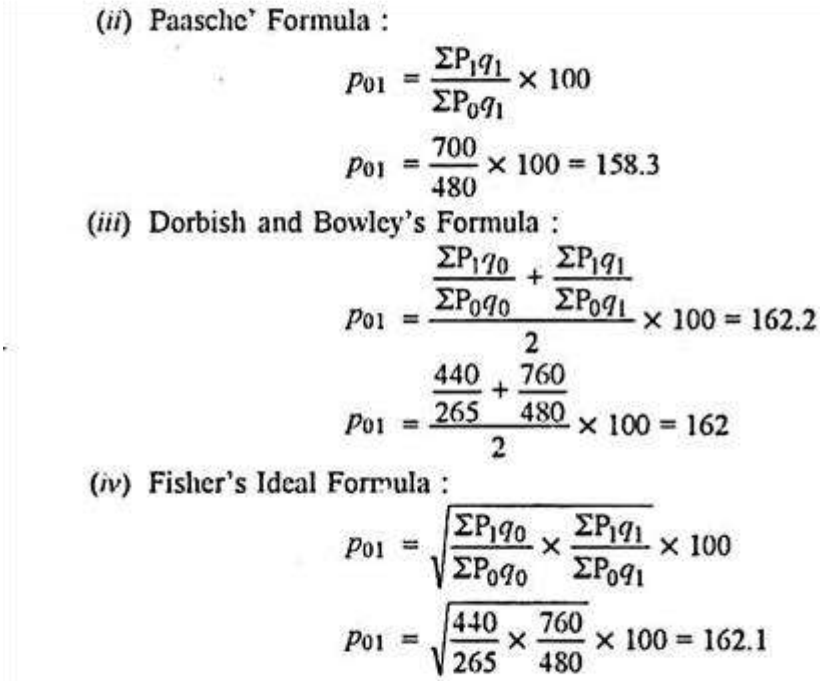
Example

Commodity	Base Year		Current Year		P_0q_0	P_1q_0	P_0q_1	P_1q_1
	P_0	q_0	P_1	q_1				
A	10	5	20	2	50	100	20	40
B	15	4	25	8	60	100	120	200
C	40	2	60	6	80	120	240	360
D	25	3	40	4	75	120	100	160
Total					265 ΣP_0q_0	440 ΣP_1q_0	480 ΣP_0q_1	760 ΣP_1q_1

(*i*) Laspeyre's Formula :

$$p_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

$$p_{01} = \frac{440}{265} \times 100 = 166.04$$



(*ii*) Paasche' Formula :

$$p_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

$$p_{01} = \frac{700}{480} \times 100 = 158.3$$

(*iii*) Dorbish and Bowley's Formula :

$$p_{01} = \frac{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 = 162.2$$

$$p_{01} = \frac{\frac{440}{265} + \frac{760}{480}}{2} \times 100 = 162$$

(*iv*) Fisher's Ideal Formula :

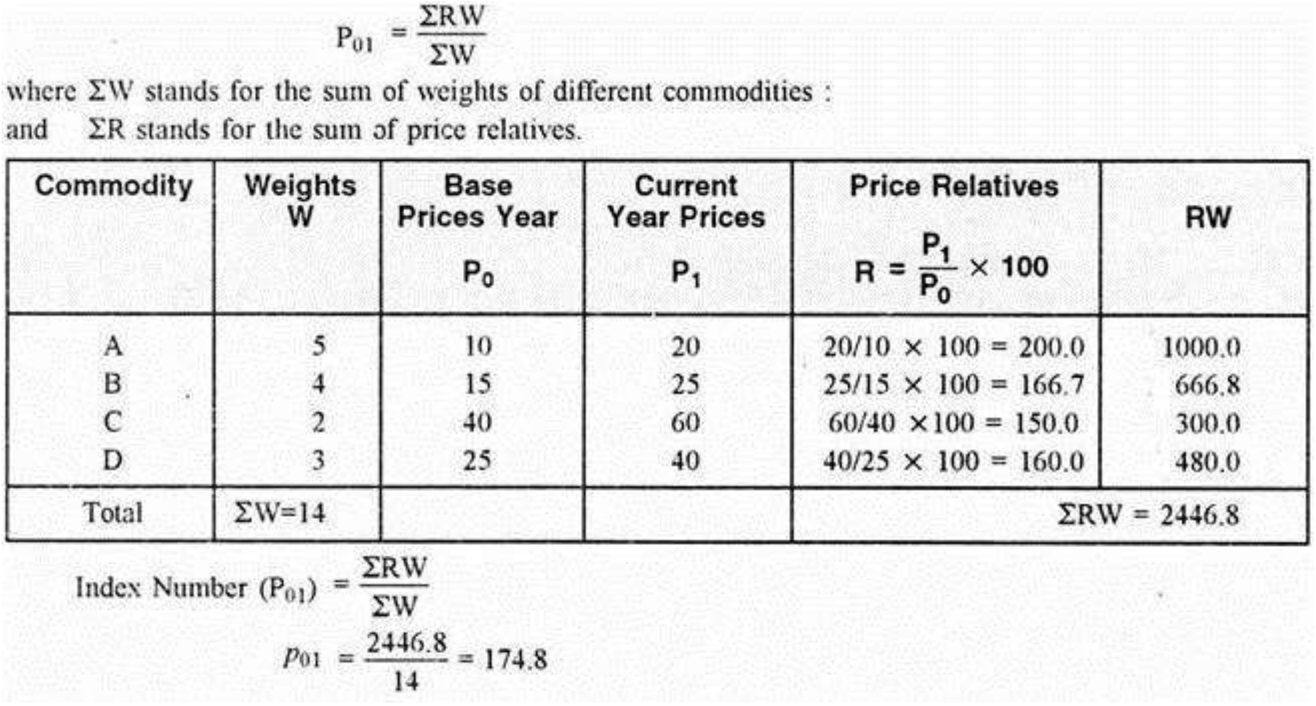
$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100$$

$$p_{01} = \sqrt{\frac{440}{265} \times \frac{760}{480}} \times 100 = 162.1$$

4. Weighted Average of Relatives Method:

In this method also different weights are used for the items according to their relative importance.

The price index number is found out with the help of the following formula:



$$P_{01} = \frac{\Sigma RW}{\Sigma W}$$

where ΣW stands for the sum of weights of different commodities :
and ΣR stands for the sum of price relatives.

Commodity	Weights W	Base Prices Year P_0	Current Year Prices P_1	Price Relatives $R = \frac{P_1}{P_0} \times 100$	RW
A	5	10	20	20/10 × 100 = 200.0	1000.0
B	4	15	25	25/15 × 100 = 166.7	666.8
C	2	40	60	60/40 ×100 = 150.0	300.0
D	3	25	40	40/25 × 100 = 160.0	480.0
Total	ΣW=14				ΣRW = 2446.8

$$\text{Index Number } (P_{01}) = \frac{\Sigma RW}{\Sigma W}$$

$$p_{01} = \frac{2446.8}{14} = 174.8$$

Quantity Index Numbers:

Now we will specifically understand what are quantity index numbers. Quantity index numbers measure the change in the quantity or volume of goods sold, consumed or produced during a given time period. Hence it is a measure of relative changes over a period of time in the quantities of a particular set of goods.

Just like price index numbers and value index numbers, there are also two types of quantity index numbers, namely

- Unweighted Quantity Indices
- Weighted Quantity Indices

Let us take a look at the various methods, formulas, and examples of both these types of quantity index numbers.

Unweighted Index: Simple Aggregate Method

Here we do a simple and direct comparison of the aggregate quantities of the current year, with those of the previous year. We express this index number as a percentage. No weights are assigned, it is the simplest calculation. The formula is as follows,

$Q01=\Sigma Q1\Sigma Q0\times 100$

where, Q1 is the quantity of the current year, and Q0 is the quantity of the previous year,

Unweighted Index: Simple Average of Quantity Method

In this method, we take the aggregate quantities of the current year as a percentage of the quantity of the base year. Then to obtain the index number, we average this percentage figure. So the formula under this method is as follows,

$Q01=\Sigma Q1\Sigma Q0\times 100\div N$

where N is the total number of items.

Weighted Index: Simple Aggregative Method:

There are a few various methods for calculating this index number. We will take a look at some of the most important ones.

1] ***Laspeyres Method:***

In this method, the base price is taken as the weight. We only use the price of the base year (P0), not the current year. The formula is as follows,

$Q01=\Sigma Q1P0\Sigma Q0P0\times 100$

2] ***Paasche's Method***

Here, the current year price (P1) of the commodity is taken as the weight.

$Q01=\Sigma Q1P0\Sigma Q0P0\times 100$

3] ***Dorbish& Bowley's Method***

$Q01=\Sigma Q1P0\Sigma Q0P0+\Sigma Q1P1\Sigma Q0P1\div 2$

Weighted Index : Weighted Average of Relative Method

In this method, we use the arithmetic mean for averaging the values. The formula is a little more complex as seen below,

$Q01=\Sigma QV\Sigma V$

where

$Q=\Sigma q1\Sigma q0$

and

$V=q0p0$

Solved Question on Quantity Index Numbers:

Q: All circumstances remaining same, a company sold 23000 tonnes of grain in the current year compared to 21500 tonnes in the last year. Find the simple unweighted quantity index number.

Ans: The formula is as follows,

$Q01=\Sigma Q1\Sigma Q0\times100$

$Q01=\Sigma23000\Sigma21500\times100=106.97$

This means the quantities saw an increase of 6.97%.

Tests of Adequacy of Index Number Formula:

- Unit Test
- Time Reversal Test
- Factor Reversal Test
- Circular Test

Unit Test:

This test states that the formula for constructing an index number should be independent of the units in which prices and quantities are expressed. All methods, except simple aggregative method, satisfy this test.

Time Reversal Test:

This test guides whether the method works both ways in time forward and backward. To quote Fisher, Time Reversal Test is the test which gives the same ratio between one point of comparison and the other, for calculation of index number, no matter which of the two is taken as base. In other words, when the index numbers of the two years are constructed by reversing the base year, they should be reciprocals of each other so that their product is unity.

Symbolically the test is represented as:

P_{01} X P_{10} = 1

where P_{01} is the index for time “1” on time “0” as base and P_{10} is the index for time “0” on time “1” as base. If the product is not unity, the method suffers from time bias.

Time reversal test is satisfied by

1. Simple aggregative method
2. Fisher’s method
3. Marshall-Edgeworth’s method and
4. Kelly’s method.

Factor Reversal Test:

Fisher has described this test in the following words: Just as each formula should permit the interchange of the two items without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent result, i.e., the two results multiplied together should give the true value ratio.”

In simple words, the test means that the change in the price multiplied by the change in the quantity should be equal to the total change in the value. The total value of a given commodity in a given year is the product of the quantity and the price per unit (total value = p x q) . The ratio of the total value in one year to the total value in the preceding

year is p_1q_1 / p_0q_0. Symbolically the test is represented as

P_{01} x Q_{01} = Σp_1q_1 / Σp_0q_0 V_{01}

Where P_{01} denotes price index and Q_{01} , quantity index number.

This test is satisfied by Fisher’s method only.

Circular Test:

According to this, if indices are constructed for year one based on year zero, for year two based on year one and for year zero based on year two, the product of all the indices should be equal to 1.

Symbolically, P_{01} X P_{12} X P_{20} = 1

This test is satisfied by

1. Simple aggregative method and
2. Kelly’s method.

Chain Index Numbers:

According to the fixed base methods, the base remains the same and unchangeable throughout the series. But, as the time passes some items may be added in the series while some may be deleted. It, therefore, becomes tough to compare the result of the current conditions with that of the past period. Thus, in such a situation changing the base period is more appropriate. Chain Index Numbers method is one such method and we shall discuss it now in detail.

Chain Index Numbers:

Under this method, firstly we express the figures for each year as a percentage of the preceding year. These are known as Link Relatives. We then need to chain them together by successive multiplication to form a chain index.

Thus, unlike fixed base methods, in this method, the base year changes every year. Hence, for the year 2001, it will be 2000, for 2002 it will be 2001, and so on. Let us now study this method step by step.

Steps in the construction of Chain Index Numbers

1. Calculate the link relatives by expressing the figures as the percentage of the preceding year. Thus,

$$\text{Link Relatives of current year} = \frac{priceofcurrentyear}{priceofpreviousyear} \times 100$$

Calculate the chain index by applying the following formula:

$$\text{Chain Index} = \frac{Currentyearrelative \times Previousyearlinkrelative}{100}$$

Advantages of Chain Index Numbers Method

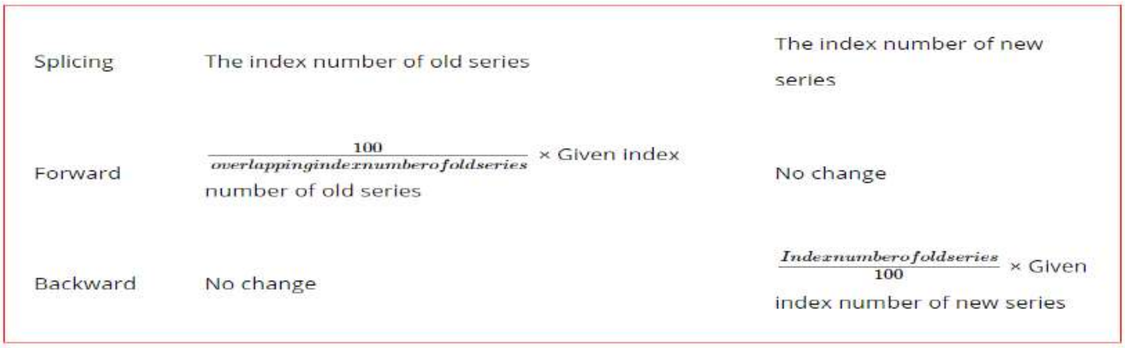
1. This method allows the addition or introduction of the new items in the series and also the deletion of obsolete items.
2. In an organization, management usually compares the current period with the period immediately preceding it rather than any other period in the past. In this method, the base year changes every year and thus it becomes more useful to the management.

Disadvantages of Chain Index Numbers Method

1. Under this method, if the data for any one of the year is not available then we cannot compute the chain index number for the subsequent period. This is so because we need to calculate the link relatives, which are not possible to be calculated in this case.
2. In case an error occurs in the calculation of any of the link relatives, then that error gets compounded and all the subsequent link relatives will also become incorrect. Thus, the entire series will give a misrepresented picture.

Splicing

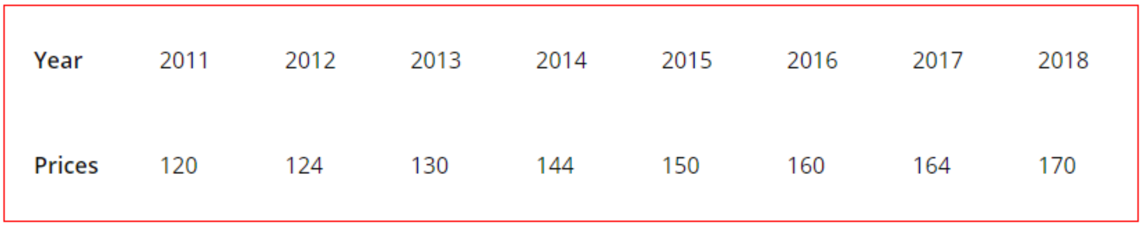
Splicing is a technique where we link the two or more index number series which contain the same items and a common overlapping year but with different base year to form a continuous series. It may be forward splicing or backward splicing. We can further understand this with the help of the table given below:



Splicing	The index number of old series	The index number of new series
Forward	$\frac{100}{overlappingindexnumberofoldseries}$ × Given index number of old series	No change
Backward	No change	$\frac{Indexnumberofoldseries}{100}$ × Given index number of new series

Solved Example on Chain Index Numbers

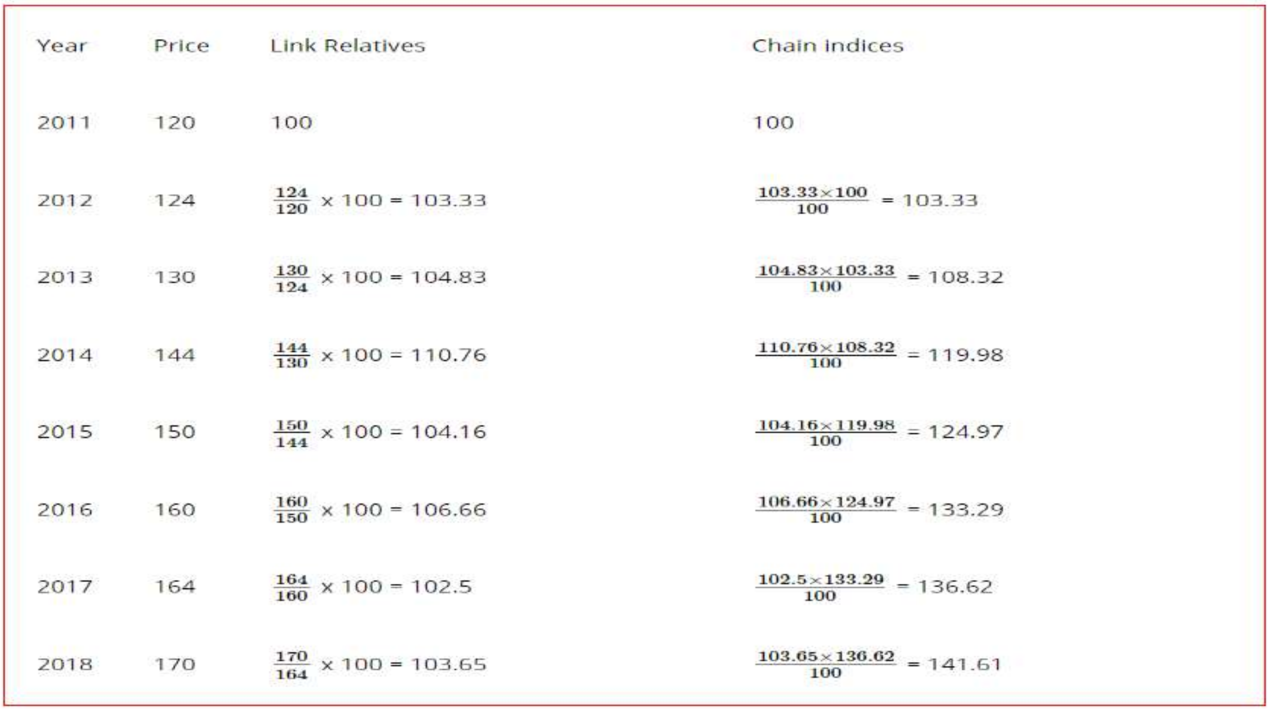
Q. From the following data calculate the index numbers using the Chain Index Numbers method.



Year	2011	2012	2013	2014	2015	2016	2017	2018
Prices	120	124	130	144	150	160	164	170

Answer:

Construction of Chain Index Numbers



Year	Price	Link Relatives	Chain indices
2011	120	100	100
2012	124	$\frac{124}{120}$ x 100 = 103.33	$\frac{103.33\times100}{100}$ = 103.33
2013	130	$\frac{130}{124}$ x 100 = 104.83	$\frac{104.83\times103.33}{100}$ = 108.32
2014	144	$\frac{144}{130}$ x 100 = 110.76	$\frac{110.76\times108.32}{100}$ = 119.98
2015	150	$\frac{150}{144}$ x 100 = 104.16	$\frac{104.16\times119.98}{100}$ = 124.97
2016	160	$\frac{160}{150}$ x 100 = 106.66	$\frac{106.66\times124.97}{100}$ = 133.29
2017	164	$\frac{164}{160}$ x 100 = 102.5	$\frac{102.5\times133.29}{100}$ = 136.62
2018	170	$\frac{170}{164}$ x 100 = 103.65	$\frac{103.65\times136.62}{100}$ = 141.61

Deflating of Index Numbers:

Deflating means making allowances for the effect of changing price levels. A rise in price level means a reduction in the purchasing power of money. The process of adjusting a series of salary or wages or income according to current price changes to find out the level of real salary wages or income is called deflating of index numbers. It is necessary when price level is increasing and cost of living is also increasing.

Consumer Price Index (CPI):

The Consumer Price Index (CPI) is a measure of the aggregate price level in an economy. The CPI consists of a bundle of commonly purchased goods and services. The CPI measures the changes in the purchasing power of a country's currency, and the price level of a basket of goods and services.

The market basket used to compute the Consumer Price Index is representative of the consumption expenditure within the economy and is the weighted average of the prices of goods and services.

Computing the Consumer Price Index

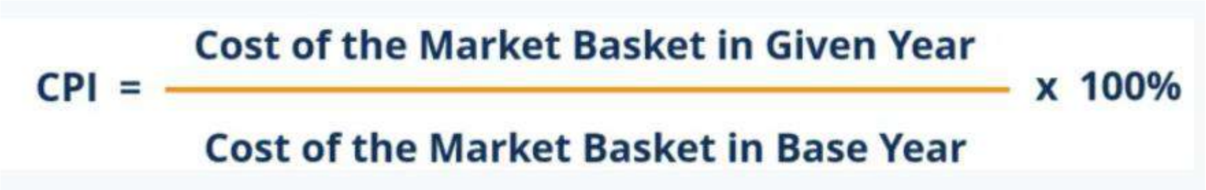
The Consumer Price Index expresses the change in the current prices of the market basket of goods in a period compared to a base period. The CPI is usually computed monthly or quarterly. It is based on a representative expenditure pattern of urban residents and includes people of all ages.

Most CPI index series use 1982-84 as the basis for comparison. The U.S. Bureau of Labor Statistics (BLS) set the index level during the 1982-84 period at 100. An index of 110 means that there's been a 10% rise in the price of the market basket compared to the reference period. Similarly, an index of 90 indicates a 10% decrease in the price of the market basket compared to the reference period.

Calculating the Consumer Price Index

The BLS records around 80,000 items each month by contacting retailers, service establishments, rental spaces, and service providers across the country.

Based on the BLS survey, the CPI is calculated using the following formula:



$$\text{CPI} = \frac{\text{Cost of the Market Basket in Given Year}}{\text{Cost of the Market Basket in Base Year}} \text{ x } 100\%$$

Determining the Market Basket (Representative Basket)

The market basket is developed using detailed expenditure information. Governments spend considerable resources (money and time) to accurately measure expenditure information. Information sources include surveys targeted at individuals, households, and businesses.

A particular item enters the basket through the initiation process. Consider the following example that describes the initiation process for bread. A particular type of bread is chosen with a probability directly proportional to its sales figures. There are three types of breads: A, B, and C. A makes up 70% of the bread market, B makes up 20% of the bread market, and C makes up 10% of the bread market.

Therefore, the probability of bread A being chosen as the representative bread is 70%. After a representative bread is chosen, its price is monitored for the next four years, after which a new representative bread will be chosen. This bread will continue to be priced each month in the same store.

Uses of the Consumer Price Index

- **To serve as an economic indicator**: The Consumer Price Index is a measure of the inflation faced by the end user. It can determine the purchasing power of the dollar. It is also a proxy for the effectiveness of a governments economic policy
- **To adjust other economic indicators** for price changes: For example, components of national income could be adjusted using CPI.
- **Provides cost of living adjustments** for wage earners and social security beneficiaries and prevents an inflation-induced increase in tax rates.

Limitations of the Consumer Price Index

- The Consumer Price Index may not be applicable to all population groups. For example, CPI-U (Urban) better represents the U.S. urban population but doesn't reflect the status of the population in rural areas.
- CPI doesn't produce official estimates for subgroups of a population.
- CPI is a conditional cost-of-living measure and does not measure every aspect that affects living standards.
- Two areas can't be compared. A higher index in one area compared to the other doesn't always mean that prices are higher in that area.
- Social and environmental factors are beyond the definitional scope of the index.

Limitations in Measurement of the CPI

- **Sampling error**: Risk of the right sample not being chosen. The sample chosen might not accurately represent the entire population.

- **Non-sampling error**: Non-sampling errors include errors associated with price-data collection and errors associated with operational implementation.
- **Doesn't include energy costs**: A major criticism of the CPI is that it doesn't include energy costs even though these are a major expenditure for most households.

Statistical quality control:

Statistical quality control, the use of statistical methods in the monitoring and maintaining of the quality of products and services. One method, referred to as acceptance sampling, can be used when a decision must be made to accept or reject a group of parts or items based on the quality found in a sample. A second method, referred to as statistical process control, uses graphical displays known as control charts to determine whether a process should be continued or should be adjusted to achieve the desired quality.

Acceptance sampling

Assume that a consumer receives a shipment of parts, called a lot, from a producer. A sample of parts will be taken and the number of defective items counted. If the number of defective items is low, the entire lot will be accepted. If the number of defective items is high, the entire lot will be rejected. Correct decisions correspond to accepting a good-quality lot and rejecting a poor-quality lot. Because sampling is being used, the probabilities of erroneous decisions need to be considered. The error of rejecting a good-quality lot creates a problem for the producer; the probability of this error is called the producer's risk. On the other hand, the error of accepting a poor-quality lot creates a problem for the purchaser or consumer; the probability of this error is called the consumer's risk.

The design of an acceptance sampling plan consists of determining a sample size *n* and an acceptance criterion *c*, where *c* is the maximum number of defective items that can be found in the sample and the lot still be accepted. The key to understanding both the producer's risk and the consumer's risk is to assume that a lot has some known percentage of defective items and compute the probability of accepting the lot for a given sampling plan. By varying the assumed percentage of defective items in a lot, several different sampling plans can be evaluated and a sampling plan selected such that both the producer's and consumer's risks are reasonably low.

Statistical Quality control

Statistical Quality control uses sampling and statistical methods to monitor the quality of an ongoing process such as a production operation. A graphical display referred to as a control chart provides a basis for deciding whether the variation in the output of a process is due to common causes (randomly occurring variations) or due to out-of-the-ordinary assignable causes. Whenever assignable causes are identified, a decision can be made to adjust the process in order to bring the output back to acceptable quality levels.

Control charts can be classified by the type of data they contain. For instance, an $\bar{x}$-chart is employed in situations where a sample mean is used to measure the quality of the output.

Quantitative data such as length, weight, and temperature can be monitored with an $\bar{x}$-chart. Process variability can be monitored using a range or *R*-chart. In cases in which the quality of output is measured in terms of the number of defectives or the proportion of defectives in the sample, an *np*-chart or a *p*-chart can be used.

All control charts are constructed in a similar fashion. For example, the centre line of an $\bar{x}$-chart corresponds to the mean of the process when the process is in control and producing output of acceptable quality. The vertical axis of the control chart identifies the scale of measurement for the variable of interest. The upper horizontal line of the control chart, referred to as the upper control limit, and the lower horizontal line, referred to as the lower control limit, are chosen so that when the process is in control there will be a high probability that the value of a sample mean will fall between the two control limits. Standard practice is to set the control limits at three standard deviations above and below the process mean. The process can be sampled periodically. As each sample is selected, the value of the sample mean is plotted on the control chart. If the value of a sample mean is within the control limits, the process can be continued under the assumption that the quality standards are being maintained. If the value of the sample mean is outside the control limits, an out-of-control conclusion points to the need for corrective action in order to return the process to acceptable quality levels.

*****