SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

Syllabus with Course objectives and Course outcomes

UNIT 1 INTRODUCTION

Introduction - Classification and Tabulation of Statistical Data - Diagrammatic and Graphical representation of data.

UNIT 2 DESCRIPTIVE STATISTICS

Measures of Central Tendency - Mean, Median and Mode - Dispersion, Range, Quartile Deviation, Mean Deviation Standard Deviation - Measures of Skewness.

UNIT 3 CORRELATION AND REGRESSION

Correlation - Karl Pearson's Coefficient of Correlation - Spearman's Rank Correlation - Regression Lines and Regression Coefficients.

UNIT 4 PROBABILITY

Definitions - Examples of Sample Space, Events, Independent Events and Conditional Events. Axiomatic and classical approach to probability. - Addition, Multiplication and Baye's theorem with Simple Applications.

UNIT 5 TIME SERIES

Time Series Analysis - Trend - Seasonal Variation

Course Objectives:

- To know the different sources and methods of data collection and different methods of data presentation.
- To understand the significance of advanced concepts of Statistics.
- To apply correlation and regression to real life problems.

Course Outcomes:

On completion of the course, the student will be able to:

CO1 Understand the meaning, scope and importance of essential concepts of Statistics

CO2 Know the different sources and method of data collection and data presentation

CO3 Measuring the central tendencies, measures of dispersion and measures of skewness.

CO4 Identify the significance of advanced concepts of Statistics

CO5 Analyse the methods of correlation and regression

CO6 Acquaint knowledge to estimate the least square methods of trend analysis

UNIT – I – Introduction to Statistics – SMTA1104

Course Material (B.Com)

Subject Name: Business Statistics

Unıt 1

INTRODUCTION

1.1 Introduction to Statistics

The word 'statistics' has been derived from the Latin word 'status'. In the plural sense it means a set of numerical figures called 'data' obtained by counting, or, measurement. In the singular sense it means collection, classification, presentation, analysis, comparison and meaningful interpretation of 'raw data'.

It has been defined in different ways by different authors. Croxton and Cowdon defined it as 'the science which deals with the collection, analysis and interpretation of numerical data'. Statistical data help us to understand the economic problems, e.g., balance of trade, disparities of income and wealth, national income accounts, supply and demand curves, living and whole sale price index numbers, production, consumption, etc., formulate economic theories and test old hypothesis. It also helps in planning and forecasting.

The success of modern business firms depends on the proper analysis of statistical data. Before expansion and diversification of the existing business or setting up a new venture, the top executives must analyze all facts like raw material prices, consumer-preferences, sales records, demand of products, labor conditions, taxes, etc., statistically. It helps to determine the location and size of business, introduce new products or drop an existing product and in fixing product price and administration. It has also wide application in Operations Research.

1.2. Functions of Statistics

(i) It simplifies complex data and helps us to study the trends and relationships of different phenomena and compare them.
(ii) It helps us to classify numerical data, measure uncertainty, test the hypothesis, formulate policies and take valid inferences.

1.3. Limitations of Statistics and its Characteristics

(i) Statistics studies a group but not an individual.
(ii) Statistics cannot be applied to study the qualitative phenomenon.
(iii) Statistical decisions are true on an average only. For better results a large number of observations are required.
(iv) Statistical data are not mathematically accurate.
(v) Statistical data must be analyzed by statistical experts otherwise the results may be misleading.

(vi) The laws of statistics are not exact like the laws of sciences.
(vii) Statistics collected for a given purpose must be used for that purpose only.
(viii) Statistical relations do not always establish the 'cause and effect" relationship between phenomena.

A statistical enquiry has four phases, namely,
(a) Collection of data
(b) Classification and tabulation of data
(c) Analysis of data
(d) Interpretation of data

1.4 Collection of Data

Data means information. Data collected expressly for a specific purpose are called 'Primary data' e.g., data collected by a particular person or organization from the primary source for his own use, collection of data about the population by censuses and surveys, etc. Data collected and published by one organization and subsequently used by other organizations are called 'Secondary data'. The various sources of collection for secondary data are: newspapers and periodicals; publications of trade associations; research papers published by university departments, U.G.C. or research bureau's; official publications of central, state and the local and foreign governments, etc. The collection expenses of primary data are more than secondary data. Secondary data should be used with care. The various methods of collection of primary data are: (i) Direct personal investigation (interview/observation); (ii) Indirect oral investigation; (iii) Data from local agents and correspondents; (iv) Mailed questionnaires; (v) Questionnaires to be filled in by enumerators; (vi) Results of experiments, etc. Data collected in this manner are called 'raw data'. These are generally voluminous and have to be arranged properly before use.

1.5. Classification of Data

Connor defined classification as: "the process of arranging things in groups or classes according to their resemblances and affinities and gives expression to the unity of attributes that may subsist amongst a diversity of individuals". The raw data, collected in real situations and arranged haphazardly, do not give a clear picture.

Thus to locate similarities and reduce mental strain we resort to classification. Classification condenses the data by dropping out unnecessary details. It facilitates comparison between different sets of data clearly showing the different points of agreement and disagreement. It enables us to study the relationship between several characteristics and make further statistical treatment like tabulation, etc. During population census, people in the country are classified according to sex (males/ females), marital status (married/unmarried), place of residence (rural/urban), Age (0-5 years, 6-10 years, 11-15 years, etc.), profession (agriculture, production, commerce, transport, doctor, others), residence in states (West Bengal, Bihar, Mumbai, Delhi, etc.), etc.

1.5.1. Modes of Classification

There are four types of classification, viz.,

(i) Qualitative
(ii) Quantitative
(iii) Geographical
(iv) Chronological

(i) Qualitative classification:

It is done according to attributes or non-measurable characteristics, like social status, sex, nationality, occupation, etc. For example, the population of the whole country can be classified into four categories as married, unmarried, widowed and divorced. When only one attribute, e.g., sex, is used for classification, it is called simple classification. When more than one attributes, e.g., deafness, sex and religion, are used for classification, it is called manifold classification.

(ii) Quantitative classification:

It is done according to numerical size like weights in kg or heights in cm. Here we classify the data by assigning arbitrary limits known as class-limits. The quantitative phenomenon under study is called a variable. For example, the population of the whole country may be classified according to different variables like age, income, wage, price, etc. Hence this classification is often called 'classification by variables'.

Variable: A variable in statistics means any measurable characteristic or quantity which can assume a range of numerical values within certain limits, e.g., income, height, age, weight, wage, price, etc. A variable can be classified as either discrete or continuous.

(1) Discrete variable: A variable which can take up only exact values and not any fractional values, is called a 'discrete' variable. Number of workmen in a factory, members of a family, students in a class, number of births in a certain year, number of telephone calls in a month, etc., are examples of discrete-variable.

(2) Continuous variable: A variable which can take up any numerical value (integral/fractional) within a certain range is called a 'continuous' variable. Height, weight, rainfall, time, temperature, etc., are examples of continuous variables. Age of students in a school is a continuous variable as it can be measured to the nearest fraction of time, i.e., years, months, days, etc

(iii) Geographical classification:

It is done according to time, e.g., index numbers arranged over a period of time, population of a country for several decades, exports and imports of India for different five year plans, etc.

(iv) Chronological classification:

It is done with respect to space or places, e.g., production of cereals in quintals in various states, population of a country according to states, etc.

1.6. Presentation of Statistical Data

Statistical data can be presented in three different ways: (a) Textual presentation, (b) Tabular presentation, and (c) Graphical presentation.

(a) Textual presentation:

This is a descriptive form. The following is an example of such a presentation of data about deaths from industrial diseases in Great Britain in 1935–39 and 1940–44.

Example 1

Numerical data with regard to industrial diseases and deaths in Great Britain during the years 1935–39 and 1940–44 are given in a descriptive form:

"During the quinquennium 1935–39, there were in Great Britain 1, 775 cases of industrial diseases made up of 677 cases of lead poisoning, 111 of other poisoning, 144 of anthrax, and 843 of gassing. The number of deaths reported was 20 p.c. of the cases for all the four diseases taken together, that for lead poisoning was 135, for other poisoning 25 and that for anthrax was 30. During the next quinquennium, 1940–44, the total number of cases reported was 2, 807. But lead poisoning cases reported fell by 351 and anthrax cases by 35. Other poisoning cases increased by 784 between the two periods. The number of deaths reported decreased by 45 for lead poisoning, but decreased only by 2 for anthrax from the pre-war to the post-war quinquennium. In the later period, 52 deaths were reported for poisoning other than lead poisoning. The total number of deaths reported in 1940–44 including those from gassing was 64 greater than in 1935–39".

The disadvantages of textual presentation are: (i) it is too lengthy; (ii) there is repetition of words; (iii) comparisons cannot be made easily; (iv) it is difficult to get an idea and take appropriate action.

(b) Tabular presentation

Tabulation may be defined as the systematic presentation of numerical data in rows or and columns according to certain characteristics. It expresses the data in concise and attractive form which can be easily understood and used to compare numerical figures.

Objectives of Tabulation:

The main objectives of tabulation are stated below:

(i) to carry out investigation; (ii) to do comparison; (iii) to locate omissions and errors in the data; (iv) to use space economically; (v) to study the trend; (vi) to simplify data; (vii) to use it as future reference.

Characteristics of a good statistical table should contain the following:

(i) Table number: A number must be allotted to the table for identification, particularly when there are many tables in a study.

(ii) Title: The title should explain what is contained in the table. It should be clear, brief and set in bold type on top of the table. It should also indicate the time and place to which the data refer.

(iii) Date: The date of preparation of the table should be given.

(iv) Stubs, or, Row designations: Each row of the table should be given a brief heading. Such designations of rows are called “stubs”, or, “stub items” and the entire column is called “stub column”.

(v) Column headings, or, Captions: Column designation is given on top of each column to explain to what the figures in the column refer. It should be clear and precise. This is called a “caption”, or, “heading”. Columns should be numbered if there are four, or, more columns.

(vi) Body of the table: The data should be arranged in such a way that any figure can be located easily. Various types of numerical variables should be arranged in an ascending order, i.e., from left to right in rows and from top to bottom in columns. Column and row totals should be given.

(vii) Unit of measurement: If the unit of measurement is uniform throughout the table, it is stated at the top right-hand corner of the table along with the title. If different rows and columns contain figures in different units, the units may be stated along with “stubs”, or, “captions”. Very large figures may be rounded up but the method of rounding should be explained.

(viii) Source: At the bottom of the table a note should be added indicating the primary and secondary sources from which data have been collected.

(ix) Footnotes and references: If any item has not been explained properly, a separate explanatory note should be added at the bottom of the table.

A table should be logical, well-balanced in length and breadth and the comparable columns should be placed side by side. Light/heavy/thick or double rulings may be used to distinguish sub columns, main columns and totals. For large data more than one table may be used.

Frequency Distribution

It is a quantitative classification of a statistical data. The following table is an example of a frequency distribution table.

Example 2

Suppose the data collected are the marks of 100 students in statistics in B.Com. course in a college. This data has to be presented in the form of a table to make further analysis and interpretation of the data. The above data can be presented in the form of a table below.

The range 0-9, 10-19, …, are called classes and the number of students in the range are called frequencies. The whole data arranged in this form is called a frequency distribution. The number of classes should not be too many. To decide the number of classes for a frequency distribution in the whole data, we choose the lowest and the highest of the values. The difference between them will enable us to decide the class intervals. The classes should be in such a way that, there is no ambiguity in deciding the class to which any value of the data should belong to. For example, if the classes are taken as 0-10, 10-20, 20-30,…, the there is ambiguity of deciding the class values 10, 20, 30, …. Such choices of classes should be avoided.

Marks Obtained	Tally Marks	No. of Students
0-9	///	3
10-19	𝍸 //	7
20-29	𝍸 𝍸 //	12
30-39	𝍸𝍸𝍸///	18
40-49	𝍸𝍸 𝍸𝍸 𝍸𝍸	25
50-59	𝍸𝍸𝍸 //	17
60-69	𝍸 ///	8
70-79	𝍸 /	6
80-89	////	4
90-99		0
Total		100

(c) Graphical presentation:

Quantitative data may also be presented graphically by using bar charts, pie diagrams, pictographs, line diagrams, etc.

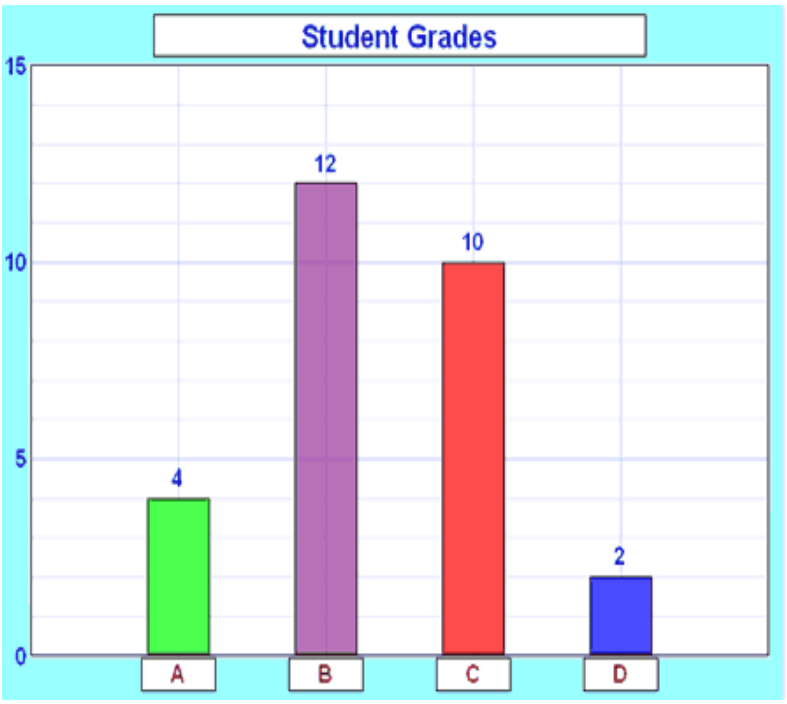
(i) Bar diagram

A bar diagram is a chart that uses bars to show comparisons between categories of data. The bars can be either horizontal or vertical. Bar graphs with vertical bars are sometimes called vertical bar graphs. A bar graph will have two axes. One axis will describe the types of categories being compared, and the other will have numerical values that represent the values of the data. It does not matter which axis is which, but it will determine what bar graph is shown. If the descriptions are on the horizontal axis, the bars will be oriented vertically, and if the values are along the horizontal axis, the bars will be oriented horizontally. There are different types of bar graphs:

(a) Simple bar diagram
(b) Multiple bar diagram
(c) Subdivided bar diagram
(d) Percentage subdivided bar diagram
(e)

Example 3 Draw a histogram for the following data.

Grade	A	B	C	D
Students	4	12	10	2



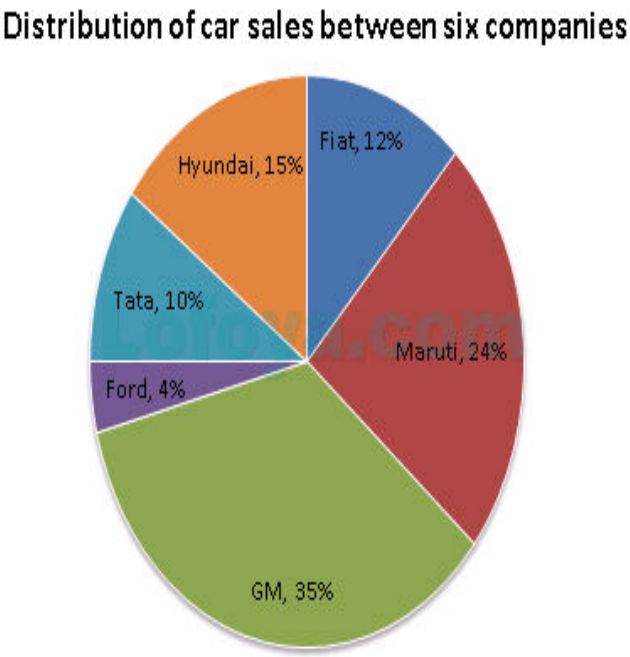
(ii) Pie diagram

Pie charts are specific types of data presentation where the data is represented in the form of a circle. In a pie chart, a circle is divided into various sections or segments such that each sector or segment represents a certain proportion or percentage of the total. The angle of a sector is proportional to the frequency of the data. The formula to determine the angle of a sector is $\frac{\text{frequency of data}}{\text{total}} \times 360°$. In such a diagram, the total of all the given items is equated to 360 degrees and the degrees of angles, representing different items, are calculated proportionately. The entire diagram looks like a pie and its components resemble slices cut from a pie. The pie chart is used to show the break-up of one continuous variable into its component parts.

For example, chart below shows the distribution of the sales of the car industry between six car companies.

Example 4

Looking at the chart below, we can infer that Maruti accounts for 24 per cent of the market share, while GM accounts for 35 percent of the market share, Ford for4 percent of the market share, Tata for 10 percent of the market share, Hyundai for 15 percent of the market share and Fiat for 12 per cent of the market share.

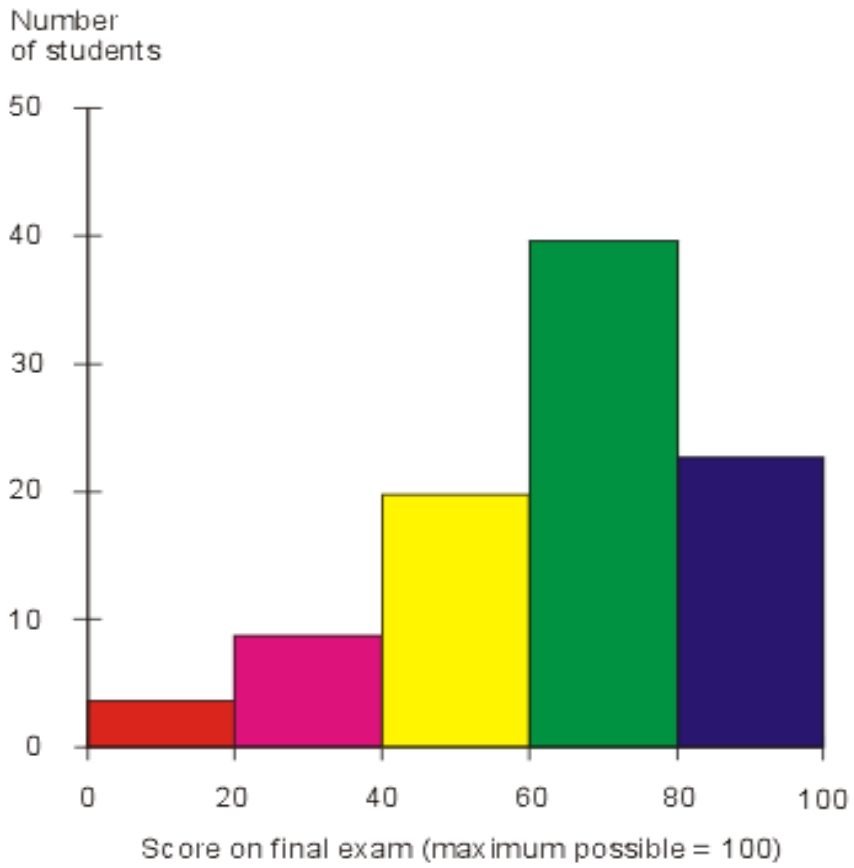


(iii) Histogram

A histogram is a display of statistical information that uses rectangles to show the frequency of data items in successive numerical intervals of equal size. In the most common form of histogram, the independent variable is plotted along the horizontal axis and the dependent variable is plotted along the vertical axis. The data appears as colored or shaded rectangles of variable area.

Example 5

The illustration, below, is a histogram showing the results of a final exam given to a hypothetical class of students. Each score range is denoted by a bar of a certain color. If this histogram were compared with those of classes from other years that received the same test from the same professor, conclusions might be drawn about intelligence changes among students over the years. Conclusions might also be drawn concerning the improvement or decline of the professor's teaching ability with the passage of time. If this histogram were compared with those of other classes in the same semester who had received the same final exam but who had taken the course from different professors, one might draw conclusions about the relative competence of the professors.



(iv) Cumulative Frequency Distribution (Ogive)

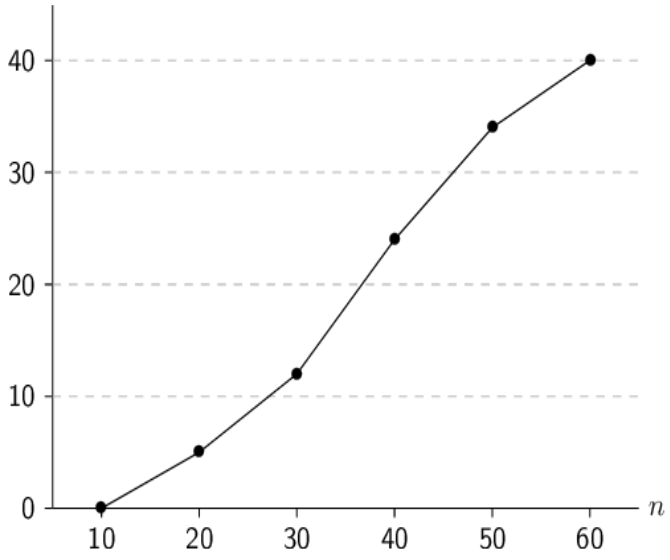
A frequency distribution becomes cumulative when the frequency of each class-interval is cumulative. Cumulative frequency of a class-interval can be obtained by adding the frequency of that class-interval to the sum of the frequencies of the preceding class-intervals. Often we want to know the number of cases which fall below, or, above a certain value. Hence, there are two types of cumulative frequencies, i.e., (1) less than (or, from below) cumulative frequency, and (2) more than (or, from above) cumulative frequencies. In the less than type the cumulative frequency of each class-interval is obtained by adding the frequencies of the given class and all the preceding classes, when the classes are arranged in the ascending order of the value of the variable. In the more than type the cumulative frequency of each class-interval is obtained by adding the frequencies of the given class and the succeeding classes. For grouped frequency distribution, the cumulative frequencies are shown against the class-boundary points.

Cumulative histograms, also known as Ogives, are graphs that can be used to determine how many data values lie above or below a particular value in a data set. The cumulative frequency is calculated from a frequency table, by adding each frequency to the total of the frequencies of all data values before it in the data set. The last value for the cumulative frequency will always be equal to the total number of data values, since all frequencies will already have been added to the previous total.

Example 6

Determine the cumulative frequencies of the following grouped data and draw an Ogive curve for the data below.

Interval	Frequency	Cumulative frequency
$10 < n \leq 20$	5	5
$20 < n \leq 30$	7	12
$30 < n \leq 40$	12	24
$40 < n \leq 50$	10	34
$50 < n \leq 60$	6	40



A less than Ogive curve

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

SMTA1104 – Business Statistics

UNIT – II – Descriptive Statistics – SMTA1104

Course Material (B.Com)

Subject Name: Business Statistics

Subject Code: SMTA1104

Unit II

Mean

The mean (also know as average), is obtained by dividing the sum of observed values by the number of observations, *n.*

Median

The median is the middle value of a set of data containing an odd number of values, or the average of the two middle values of a set of data with an even number of values.

Arrange all of the values from lowest to highest. If there are an odd number of entries, the median is the middle value. If there are an even number of entries, the median is the mean of the two middle entries.

Mode

The mode is the most frequently occurring value in the data set. In a data set where each value occurs exactly once, there is no mode.

INDIVIDUAL SERIES	**DISCRETE SERIES**	**CONTINUOUS SERIES**
ARITHMETIC MEAN: Direct Method $\bar{X} = \frac{\sum X}{N}$ Short-cut Method $\bar{X} = A + \frac{\sum d}{N}$ Step-Deviation Method $\bar{X} = A + \frac{\sum d}{N} \times i$	Direct Method $\bar{X} = \frac{\sum f X}{N}$ Short-cut Method $\bar{X} = A + \frac{\sum f d}{N}$ Step-Deviation Method $\bar{X} = A + \frac{\sum f d}{N} \times i$	Direct Method $\bar{X} = \frac{\sum f X}{N}$ Short-cut Method $\bar{X} = A + \frac{\sum f d}{N}$ Step-Deviation Method $\bar{X} = A + \frac{\sum f d}{N} \times i$
MEDIAN: Size of $\left(\frac{N+1}{2}\right)^{th}$ term	Size of $\left(\frac{N+1}{2}\right)^{th}$ term	Size of $\left(\frac{N}{2}\right)^{th}$ term Median = $L + \frac{N/2 - c.f.}{N} \times i$
MODE:		

Either by inspection or the value that occurs largest number of times	Grouping Method determines that value around which most of the frequencies are concentrated.	$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times i$
EMPIRICAL RELATION: Mode = 3 Mean – 2 Median		

EXAMPLES

1. Find the mode, median and mean for this set of numbers: 3, 6, 9, 14, 3

Solution

First arrange the numbers from least to greatest: 3, 3, 6, 9, 14

Mode (number seen most often) = 3

3, 3, 6, 9, 14

Median (number exactly in the middle) = 6

3, 3, 6, 9, 14

Mean (add up all the numbers then divide by the amount of numbers) = 7

3 + 3 + 6 + 9 + 14 = 35 35 / 5 = 7

2. Find the mode, median and mean for this set of numbers: 1, 8, 23, 7, 2, 5

Solution

First arrange the numbers from least to greatest: 1, 2, 5, 7, 8, 23

Mode = no mode

Median = 6 (5 + 7 = 12 ; 12 / 2 = 6)

Range = 22 (23 – 1 = 22)

Mean = 7 4/6 (1 + 2 + 5 + 7 + 8 + 23 = 46 ; 46 / 6 = 7 4/6 or 7 2/3. This answer may also be stated in decimals.)

3. From the following data compute Arithmetic Mean

Marks	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 – 60
No. of students	5	10	25	30	20	10

Solution:

Marks	Midvalue X	No. of students f	f x
0 – 10	5	5	25
10 – 20	15	10	150
20 – 30	25	25	625
30 – 40	35	30	1050

40 – 50	45	20	900
50 – 60	55	10	550
		N=100	3300

Arithmetic Mean $\bar{X} = \frac{\sum f X}{N} = \frac{3300}{100} = 33$

4. Calculate Arithmetic Mean from the following data

Marks	0 - 10	10 - 30	30 - 60	60 - 100
No. of students	5	12	25	8

Solution:

The class intervals are unequal but still to simplify calculations we can take 5 as common factor.

Marks	Midvalue x	No. of students F	d (x - 45) / 5	f d
0 – 10	5	5	- 8	- 40
10 – 30	15	12	- 5	- 60
30 – 60	25	25	0	0
60 - 100	35	8	7	56
		N= 50		- 44

Arithmetic Mean $\bar{X} = A + \frac{\sum f d}{N} \times i$

A = 45, $\sum f d$ = - 44, N = 50, I = 5

$$\bar{X} = 45 - \frac{44}{50} \times 5 = 45 - 4.4 = 40.6$$

5. Find the missing frequency from the following data

Marks	0 – 10	10 – 20	20 - 30	30 - 40	40 -50	50 – 60
No. of Students	5	15	20	-	20	10

The arithmetic mean is 34 marks.

Solution:

Let the missing frequency be denoted by X

Marks	Midvalue x	f	f x
0 – 10	5	5	25
10 – 20	15	15	225
20 – 30	25	20	500
30 – 40	35	X	35X
40 -50	45	20	900
50 – 60	55	10	550
		N = 70 + X	2200+35X

$$\bar{X} = \frac{\Sigma f x}{N} \qquad 34 = \frac{2200 + 35X}{70 + X}$$

$$34\ (70 + X) = 2200 + 35X$$

$$2380 + 34X = 2200 + 35X$$

$$35X - 34X = 2380 - 2200$$

$$X = 180$$

6. Calculate the Median and Mode from the following data

Central size	15	25	35	45	55	65	75	85
Frequencies	5	9	13	21	20	15	8	3

Solution:

Since we are given central values first we determine the lower and upper limits of the classes. The class interval is 10 and hence the first class would be 10 – 20.

Class	Midvalue X	f	d (x – 55) / 10	f d	c.f.
10 - 20	15	5	- 4	- 20	5
20 - 30	25	9	- 3	- 27	14
30 - 40	35	13	- 2	- 26	27
40 -50	45	21	- 1	- 21	48
50 - 60	55	20	0	0	68
60 - 70	65	15	1	15	83
70 - 80	75	8	2	16	91
80 - 90	85	3	3	9	94
				Σfd = - 54	

Calculation of Median:

Med = size of $\frac{N}{2}$ th term = $\frac{94}{2} = 47$

Median lies in the class 40 – 50

Median = $L + \frac{N/2 - c.f.}{f} \times i$

$$M = 40 + \frac{47 - 27}{21} \times 10 = 40 + 9.524 = 49.524$$

7. Calculate the median and mode of the data given below. Using then find arithmetic mean

Marks	0 – 10	10 – 20	20 - 30	30 – 40	40 -50	50 – 60
No. of Students	8	23	45	65	75	80

Solution:

Marks	f	c.f.
0 – 10	8	8
10 – 20	15	23
20 – 30	22	45
30 – 40	20	65
40 – 50	10	75
50 – 60	5	80
	N = 80	

Calculation of Median: Med = size of $\frac{N}{2}$ th term = $\frac{80}{2} = 40$ th item

Median lies in the class 20 – 30

$$\text{Median} = L + \frac{N/2 - c.f.}{f} \times i$$

$$M = 20 + \frac{40 - 23}{22} \times 10 = 20 + 7.73 = 27.73$$

Mode lies in the class is 20 – 30

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times i = 20 + \left(\frac{22 - 15}{44 - (15 + 20)} \right) \times 10 = 27.78$$

<u>MEASURES OF DISPERSION</u>: CONCEPTS AND FORMULAE

Standard deviation

Standard deviation measures the variation or dispersion exists from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread over a large range of values.

INDIVIDUAL OBERSERVATIONS	**DISCRETE& CONTINUOUS SERIES**
QUARTILE DEVIATION: Q.D. = $\frac{Q_3 - Q_1}{2}$ Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$	Quartile Deviation: Q.D. = $\frac{Q_3 - Q_1}{2}$ Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

STANDARD DEVIATION: Actual Mean Method: $\sigma = \sqrt{\frac{\sum f(X-\bar{X})^2}{N}}$ Assumed Mean Method: $\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$ Step Deviation Method $\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} \times i$ $C.V. = \frac{\sigma}{\bar{X}} \times 100$	Actual Mean Method: $\sigma = \sqrt{\frac{\sum f(X-\bar{X})^2}{N}}$ Assumed Mean Method: $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$ Step Deviation Method $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$ $C.V. = \frac{\sigma}{\bar{X}} \times 100$

EXAMPLES:

1. Find the Mean and standard deviation from the following distribution

Mid value	12.0	12.5	13.0	13.5	14	14.5	15	15.5	16
No. of Students	2	16	36	60	76	37	18	3	2

Solution:

Midvalue x	No. of Students f	d (x – 14) / 0.5	f d	f d^2
12.0	2	-4	- 8	32
12.5	16	-3	- 48	144
13	36	-2	- 72	144
13.5	60	-1	- 60	60
14	76	0	0	0
14.5	37	1	37	37
15	18	2	36	72
15.5	3	3	9	27

16.0	2	4	8	32
	N= 250		Σfd = - 98	$\Sigma fd^2 = 548$

Mean $\bar{X} = A + \frac{\Sigma f d}{N} \times i = 14 - \frac{98}{250} \times 0.5 = 13.8$

Standard deviation $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$

$$\sigma = \sqrt{\frac{548}{250} - \left(\frac{-98}{250}\right)^2} \times .05 \;=\; 0.715$$

2. Find the Standard deviation and Coefficient of Variation from the following data

Marks	No. of students
Up to 10	12
Up to 20	30
Up to 30	65
Up to 40	107
Up to 50	157
Up to 60	202
Up to 70	222
Up to 80	230

Solution:

Class	Midvalue X	No. of Students f	d (x – 35) / 10	f d	f d^2
0 – 10	5	12	-3	- 36	108
10 - 20	15	18	-2	- 36	72
20 - 30	25	35	-1	- 35	35
30 - 40	35	42	0	0	0
40 -50	45	50	1	50	50
50 - 60	55	45	2	90	180
60 - 70	65	20	3	60	180
70 - 80	75	8	4	32	128
		N= 230		Σfd =125	$\Sigma fd^2 = 753$

Mean $\bar{X} = A + \frac{\Sigma f d}{N} \times i = 35 + \frac{125}{230} \times 10 = 40.43$

Standard deviation $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$

$$\sigma = \sqrt{\frac{753}{230} - \left(\frac{125}{230}\right)^2} \times 10 \;=\; 17.26$$

$$C.V. = \frac{\sigma}{\bar{X}} \times 100 \;=\; \frac{17.26}{40.43} \times 100 = 42.69$$

3. The scores of two batsmen A and B in ten innings during a certain season are:

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Find which of the two batsmen more consistent in scoring

Solution:

X	$X-\bar{X}$	$\left(X-\bar{X}\right)^2$	Y	$Y-\bar{Y}$	$\left(Y-\bar{Y}\right)^2$
32	-14	196	19	- 31	961
28	-18	324	31	- 19	361
47	1	1	48	- 2	4
63	17	289	53	3	9
71	25	625	67	17	289
39	-7	49	90	40	1600
10	- 36	1296	10	- 40	1600
60	14	196	62	12	144
96	50	2500	40	-10	100
14	- 32	1024	80	30	900
$\sum X = 460$		$\sum\left(X-\bar{X}\right)^2 = 6500$	$\sum Y = 500$		$\sum\left(Y-\bar{Y}\right)^2 = 5968$

Batsman A:

Mean $\bar{X} = \frac{\sum X}{N} = \frac{460}{10} = 46$

$$\sigma = \sqrt{\frac{\sum(\mathrm{X}-\bar{\mathrm{X}})^2}{\mathrm{N}}} = \sqrt{\frac{6500}{10}} = 25.495$$

$$C.V. = \frac{\sigma}{\bar{X}} \times 100 = \frac{25.495}{46} \times 100 = 55.42$$

Batsman B:

Mean $\bar{Y} = \frac{\sum Y}{N} = \frac{500}{10} = 50$

$$\sigma = \sqrt{\frac{\sum\left(Y-\bar{Y}\right)^2}{N}} = \sqrt{\frac{5968}{10}} = 24.43$$

$$C.V. = \frac{\sigma}{\bar{Y}} \times 100 = \frac{24.43}{50} \times 100 = 48.86$$

Since Coefficient of Variation is less in the case of Batsman B, we conclude that the Batsman B is more consistent.

4. Calculate the Quartile deviation and the coefficient of quartile deviation from the following data

Marks	No. of students
Below 20	8
Below 40	20
Below 60	50
Below 80	70
Below 100	80

Solution:

Marks	f	c.f.
0 - 20	8	8
20 - 40	12	20
40 - 60	30	50
60 - 80	20	70
80 -100	10	80
	N= 80	

Q_1 is the size of $N / 4^{th}$ item.
Q_1 lies in the class 20 – 40

$$Q_1 = L + \frac{N/4 - c.f.}{f} \times i = 20 + \frac{20-8}{12} \times 20 = 40$$

Q_3 is the size of $3N / 4^{th}$ item.
Q_3 lies in the class 60 – 80

$$Q_3 = L + \frac{3N/4 - c.f.}{f} \times i = 60 + \frac{60-50}{20} \times 20 = 70$$

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{70-40}{2} = 15$$

$$Coefficient\ of\ QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{30}{110} = 0.273$$

5. Calculate the Inter-Quartile range and the coefficient of quartile deviation from the following data

Marks	No. of students
Above 0	150
Above 10	140
Above 20	100

Above 30	80
Above 40	80
Above 50	70
Above 60	30
Above 70	14
Above 80	0

Solution:

Marks	f	c.f.
0 - 10	10	10
10 - 20	40	50
20 - 30	20	70
30 - 40	0	70
40 -50	10	80
50-60	40	120
60-70	16	136
70-80	14	150
	N = 150	

Q_1 is the size of $N/4^{th}$ item. Q_1 lies in the class 10 – 20

$$Q_1 = L + \frac{N/4 - c.f.}{f} \times i = 10 + \frac{37.5 - 10}{40} \times 10 = 16.875$$

Q_3 is the size of $3N/4^{th}$ item. Q_3 lies in the class 50 – 60

$$Q_3 = L + \frac{3N/4 - c.f.}{f} \times i = 50 + \frac{112.5 - 80}{40} \times 10 = 58.25$$

$$Inter\ Quartile\ Range = Q_3 - Q_1 = 41.375$$

$$Coefficient\ of\ QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{41.375}{75} = 0.55$$

<u>MOMENTS</u>: FORMULAE

Moments about mean $\mu_1 = \frac{\sum(X - \overline{X})}{N} = 0$ $\mu_2 = \frac{\sum(X - \overline{X})^2}{N}$	 $\mu_3 = \frac{\sum(X - \overline{X})^3}{N}$ $\mu_4 = \frac{\sum(X - \overline{X})^4}{N}$
In a Frequency distributiom	

$\mu_1 = \frac{\sum f\,(X - \overline{X})}{N} = 0$ $\mu_2 = \frac{\sum f\,(X - \overline{X})^2}{N}$	$\mu_3 = \frac{\sum f\,(X - \overline{X})^3}{N}$ $\mu_4 = \frac{\sum f\,(X - \overline{X})^4}{N}$
Moments about arbitrary origin $\mu_1' = \frac{\sum (X - A)}{N}$ $\mu_2' = \frac{\sum (X - A)^2}{N}$	$\mu_3' = \frac{\sum (X - A)^3}{N}$ $\mu_4' = \frac{\sum (X - A)^4}{N}$
In a frequency distribution $\mu_1' = \frac{\sum f\,(X - A)}{N}$ $\mu_2' = \frac{\sum f\,(X - A)^2}{N}$	$\mu_3' = \frac{\sum f\,(X - A)^3}{N}$ $\mu_4' = \frac{\sum f\,(X - A)^4}{N}$
Moments about mean $\mu_2 = \mu_2' - \mu_1'^2$, $\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3$	$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

SKEWNESS AND KURTOSIS

Karl Pearson's Skewness = Mean - Mode
Bowley's Skewness = $Q_3 + Q_1 - 2$ Med
Karl Pearson's coefficient of Skewness = $\frac{Mean - Mode}{\sigma}$
Bowley's coefficient of Skewness = $\frac{Q_3 + Q_1 - 2\,Med}{Q_3 - Q_1}$
$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \beta_2 = \frac{\mu_4}{\mu_2^2}$
$\gamma_1 = \frac{\mu_3}{\mu_3^{\frac{3}{2}}}, \gamma_2 = \beta_2 - 3$

1. Calculate the coefficient of skewness by Karl Pearson's method and the values of β_1 and β_2 from the following data

Profits (in lakhs)	No. of companies
10 - 20	18
20 - 30	20
30 - 40	30
40 – 50	22
50 – 60	10

Solution :

Class	Midvalue x	No. of Students f	d (x – 35) / 10	f d	$f d^2$	$f d^3$	$f d^4$
10 - 20	15	18	-2	- 36	72	-144	288
20 - 30	25	20	-1	- 20	35	- 20	20
30 - 40	35	30	0	0	0	0	0
40 -50	45	22	1	22	50	22	22
50 - 60	55	10	2	20	180	80	160
		N= 100		$\Sigma fd = -14$	$\Sigma fd^2 = 154$	$\Sigma fd^3 = -62$	$\Sigma fd^4 = 490$

$$\overline{X} = A + \frac{\Sigma f\, d}{N} \times i = 35 - \frac{14}{100} \times 10 = 33.6$$

Modal class 30 -40

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times i = 30 + \frac{30 - 20}{60 - (20 + 22)} = 35.56$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$$= \sqrt{\frac{154}{100} - \left(\frac{-14}{100}\right)^2} \times 10 = 12.33$$

$$\text{Karl Pearson's coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$= \frac{33.6 - 35.56}{12.33} = -0.159$$

$$\mu_1' = \frac{\Sigma fd}{N} = -0.14$$

$$\mu_2' = \frac{\Sigma f\, d^2}{N} = 1.54$$

$$\mu_3' = \frac{\sum f\, d^3}{N} = -0.62$$

$$\mu_4' = \frac{\sum f\, d^4}{N} = 4.9$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 1.5204,$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = 0.0213$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 = 4.735$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0.00045}{3.51458} = 0.000128$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{4.735}{2.312} = 2.048$$

2. By measuring the quartiles find a measure of skewness for the following distribution

Annual Sales	No. of firms
Less than 20	30
Less than 30	225
Less than 40	465
Less than 50	580
Less than 60	634
Less than 70	644
Less than 80	650
Less than 90	665
Less than 100	680

Solution:

Sales	f`	c.f.
10 - 20	30	30
20 - 30	195	225
30 - 40	240	465
40 - 50	115	580
50 - 60	54	634
60 - 70	10	644
70 - 80	6	650
80 - 90	15	665
90 - 100	15	680
	N = 680	

Q_1 lies in the class 20-30

$$Q_1 = L + \frac{N/4 - c.f.}{f} \times i = 20 + \frac{170 - 30}{195} \times 10 = 27.18$$

. Q_3 lies in the class 40 – 50

$$Q_3 = L + \frac{3N/4 - c.f.}{f} \times i = 40 + \frac{510 - 465}{115} \times 10 = 43.9$$

$$\text{Inter Quartile Range} = Q_3 - Q_1 = 41.375$$

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{41.375}{75} = 0.55$$

Median class 30 - 40

$$\text{Median} = L + \frac{N/2 - c.f.}{f} \times i$$

$$= 30 + \frac{340 - 225}{240} \times 10 = 34.79$$

$$\text{Bowley's coefficient of Skewness} = \frac{Q_3 + Q_1 - 2\,\text{Med}}{Q_3 - Q_1}$$

$$= \frac{43.9 + 27.18 - 2(34.79)}{43.9 - 27.18} = 0.09$$

3. Calculate the first four moments about the mean from the following data and also calculate the values of β_1 and β_2

Marks	No. of students
0 - 10	5
10 - 20	12
20 - 30	18
30 - 40	40
40 - 50	15
50 - 60	7
60 - 70	3

Solution :

Class	Midvalue x	No. of Students f	d (x – 35) / 10	f d	f d^2	f d^3	f d^4
0 - 10	5	5	-3	-15	45	-135	405
10 - 20	15	12	-2	-24	48	- 96	192
20 - 30	25	18	-1	-18	18	- 18	18
30 - 40	35	40	0	0	0	0	0
40 - 50	45	15	1	15	15	15	15

50 - 60	55	7	2	14	28	56	112
60 - 70	65	3	3	9	27	81	243
		N= 100		$\Sigma f\, d = 19$	$\Sigma f\, d^2 = 181$	$\Sigma f\, d^3 = 97$	$\Sigma f\, d^4 = 985$

$$\mu_1' = \frac{\Sigma fd}{N} \times i = -1.9$$

$$\mu_2' = \frac{\Sigma f\, d^2}{N} \times i^2 = 181$$

$$\mu_3' = \frac{\Sigma f\, d^3}{N} \times i^3 = -970$$

$$\mu_4' = \frac{\Sigma f\, d^4}{N} \times i^4 = 98500$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 177.39,$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = 47.982$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'\left(\mu_1'\right)^2 - 3\left(\mu_1'\right)^4 = 95009.364$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{2302.27}{5581968.75} = 0.0004$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{95009.364}{31467.212} = 3.02$$

4. The first four moments of a distribution of a distribution about x = 2 are -2, 12, -20 and 100. Calculate the moment about mean. Also calculate β_2 and find whether the distribution is leptokurtic or platykurtic.

Solution:

$\mu_1' = -2$, $\mu_2' = 12$, $\mu_3' = -20$, $\mu_4' = 100$

$$\mu_2 = \mu_2' - \mu_1'^2 = 8,$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = 36$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'\left(\mu_1'\right)^2 - 3\left(\mu_1'\right)^4 = 20$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} == 0.3125$$

Since β_2 is less than 3 the distribution is platykurtic.

EXERCISE PROBLEMS:

1. Define measure of central tendency.

2. State Karl Pearson's coefficient of skewness.

3. Find the mode of: 4, 8, 3, 8, 8, 9, 1, 8, 3.

4. Define Kurtosis and write the measures of kurtosis.

5. Compute the median for the following frequency distribution

Class: 0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
F : 32	65	100	184	288	167	98	46	20

6. For a group of 200 candidates, the mean and standard deviation of scores were found to be 40 and 15 respectively. Later on it was discovered that the scores were found to be 43 and 35 were missed as 34 and 53 respectively. Find the correlated mean and standard deviation corresponding to the corrected figures.

7. Calculate the Arithmetic M can of the following frequency distribution:

x	0-10	10-20	20-30	30-40	40-50	50-60
f	12	18	27	20	17	6

8. In ten cricket matches two batsmen A and B scored as follows

A	12	115	6	73	7	19	119	36	84	29
B	47	12	16	42	4	51	37	48	13	0

Who is better scorer and who is more consistent

9. Calculate the coefficient of skewness and kurtosis on the moments for the following distribution

x	4.5	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5	94.5
f	1	5	12	22	17	9	4	3	1	1

10. The marks obtained by 12 students out of 50 are:

25, 20, 23, 32, 40, 27, 30, 25, 20, 10, 15, 14. Find the mean of the marks.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

SMTA1104 – Business Statistics

UNIT – III – Correlation and Regression – SMTA1104

COURES METERIAL

I B.Com

SUBJECT: BUSINESS STATISTICS

UNIT III - Correlation and Regression

Contents

Introduction

There are situations where data appears as pairs of figures relating to two variables. A correlation problem considers the joint variation of two measurements neither of which is restricted by the experimenter. The regression problem discussed in this Lesson considers the frequency distribution of one variable (called the dependent variable) when another (independent variable) is held fixed at each of several levels.

Examples of correlation problems are found in the study of the relationship between IQ and aggregate percentage of marks obtained by a person in the SSC examination, blood pressure and metabolism or the relation between height and weight of individuals. In these examples both variables are observed as they naturally occur, since neither variable is fixed at predetermined levels.

Examples of regression problems can be found in the study of the yields of crops grown with different amount of fertilizer, the length of life of certain animals exposed to different levels of radiation, and so

on. In these problems the variation in one measurement is studied for particular levels of the other variable selected by the experimenter.

Correlation

Correlation measures the degree of linear relation between the variables. The existence of correlation between variables does not necessarily mean that one is the cause of the change in the other. It should noted that the correlation analysis merely helps in determining the degree of association between two variables, but it does not tell anything about the cause and effect relationship. While interpreting the correlation coefficient, it is necessary to see whether there is any cause and effect relationship between variables under study. If there is no such relationship, the observed is meaningless. In correlation analysis, all variables are assumed to be random variables

The Scatter Diagram

The first step in correlation and regression analysis is to visualize the relationship between the variables. A scatter diagram is obtained by plotting the points (x1, y1), (x2, y2), …, (xn,yn) on a two-dimensional plane. If the points are scattered around a straight line , we may infer that there exist a linear relationship between the variables. If the points are clustered around a straight line with negative slope, then there exist negative correlation or the variables are inversely related (i.e, when x increases y decreases and vice versa.). If the points are clustered around a straight line with positive slope, then there exist positive correlation or the variables are directly related (i.e, when x increases y also increases and vice versa.).

For example, we may have figures on advertisement expenditure (X) and Sales (Y) of a firm for the last ten years, as shown in Table 1. When this data is plotted on a graph as in Figure 1 we obtain a scatter diagram. A scatter diagram gives two very useful types of information. First, we can observe patterns between variables that indicate whether the variables are related. Secondly, if the variables are related we can get an idea of what kind of relationship (linear or non-linear) would describe the relationship.

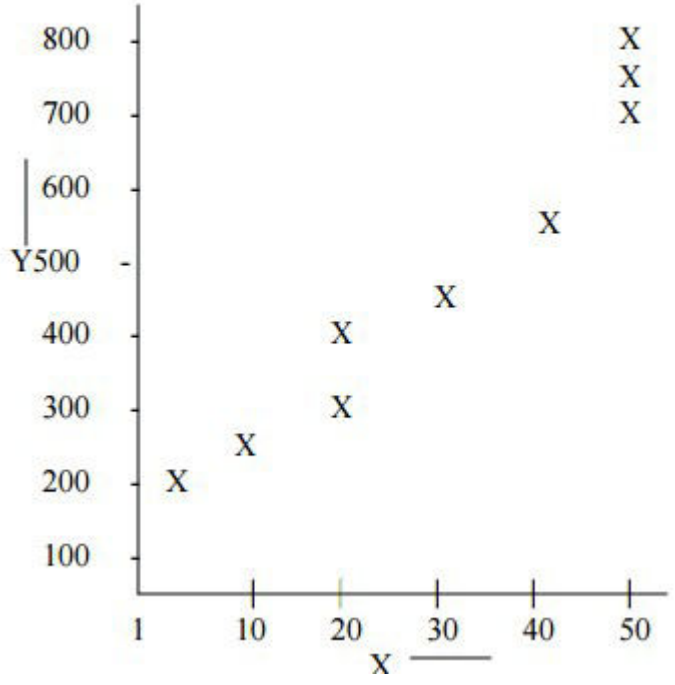
Table 1

Year-wise data on Advertisement Expenditure and Sales

Year	Advertisement Expenditure In thousand Rs. (X)	Sales in Thousand Rs. (Y)
1988	50	700
1987	50	650
1986	50	600
1985	40	500
1984	30	450
1983	20	400
1982	20	300
1981	15	250
1980	10	210
1979	5	200

Correlation examines the first Question of determining whether an association exists between the two variables, and if it does, to what extent. Regression examines the second question of establishing an appropriate relation between the variables

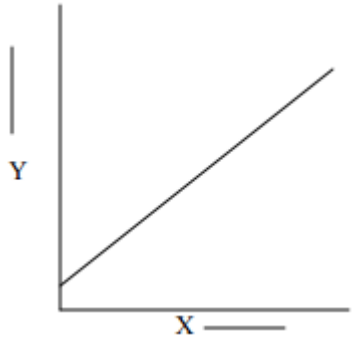
Figure 1 : Scatter Diagram



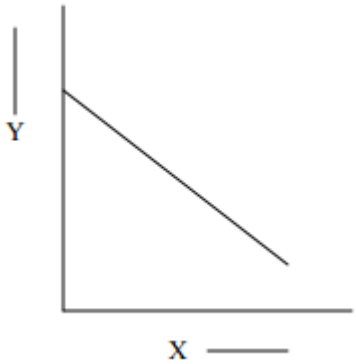
The scatter diagram may exhibit different kinds of patterns. Some typical patterns indicating different correlations between two variables are shown in Figure 2.

Figure 2: Different Types of Association between Variables

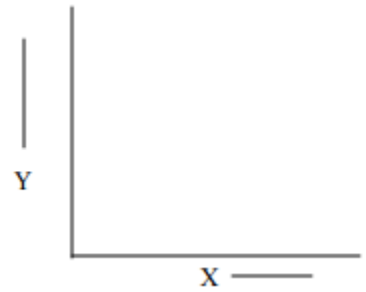
r>0 (a) Positive correlation



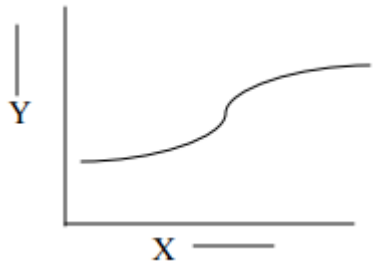
r<0 (b) negative correlation



r = 0 (c) No correlation



(d) Non-linear Association



The Correlation Coefficient

Definition and Interpretation

The correlation coefficient measure the degree of association between two variables X and Y. Pearson's formula for correlation coefficient is given as

$$r = \frac{\frac{1}{n}\sum (X - \overline{X})\,(Y - \overline{Y})}{\sigma x \sigma y}$$

Where r is the correlation coefficient between X and Y, $\sigma x and \sigma y$ are the standard deviation of X and Y respectively and n is the number of values of the pair of variables X and Y in the given data.

The expression

$$\frac{1}{n}\sum (X - \overline{X}) \quad (X - \overline{Y})$$

is known as the covariance between X and Y. Here r is also called the Pearson's product moment correlation coefficient. You should note that r is a dimensionless number whose numerical value lies between +1 and -1. Positive values of r indicate positive (or direct) correlation between the two variables

X and Y i.e. as X increase Y will also increase or as X decreases Y will also decrease. Negative values of r indicate negative (or inverse) correlation, thereby meaning that an increase in one variable results in a decrease in the value of the other variable. A zero correlation means that there is a o association between the two variables.

Figure 2 shown a number of scatter plots with corresponding values for the correlation coefficient r. The following form for carrying out computations of the correlation coefficient is perhaps more convenient:

$$r = \frac{\sum xy}{\sqrt{\sum X^2}\sqrt{\sum y^2}}$$

Where

$x = X - \overline{X}$ = deviation of a particular X value from the mean- $\overline{X}$
$y = Y - \overline{Y}$ = deviation of a particular Y value from the mean $\overline{Y}$

$$\sigma x = \sqrt{\frac{1}{n}\sum (X - \overline{X})^2} \quad \textbf{and} \quad \sigma y = \sqrt{\frac{1}{n}\sum (X - \overline{Y})^2}$$

Karl Pearson's Correlation Coefficient(Product moment correlation)

If $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n given observations, then the Karl Pearson's correlation coefficient is defined as, $r = \frac{S_{xy}}{S_x S_y}$, where S_{xy} is the covariance and S_x, S_y are the standard deviations of X and Y respectively.

That is, $$r = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2}\sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2}}$$

The value of r is in in between –1 and 1. That is, $-1 \le r \le 1$. When r = 1, there exist a perfect positive linear relation between x and y. when r = -1, there exist perfect negative linear relationship between x and y. when r = 0, there is no linear relationship between x and y.

Correlation coefficient is also represented as

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}}$$

Relation between Regression Coefficients and Correlation Coefficient

Correlation coefficient is the geometric mean of the regression coefficients.

We know that $b_{yx} = \frac{S_{xy}}{S_x^2}$ and $b_{xy} = \frac{S_{xy}}{S_y^2}$

The geometric mean of b_{yx} and b_{xy} is $\sqrt{b_{xy}b_{yx}} = \sqrt{\frac{S_{xy}S_{xy}}{S_y^2 S_x^2}}$

$$= \frac{S_{xy}}{S_x S_y}$$

= r, the correlation coefficient.

Also note that the sign of both the regression coefficients will be same, so the sign of correlation coefficient is same as the sign of regression coefficient

Coefficient of Determination

Coefficient of determination is the square of correlation coefficient and which gives the proportion of variation in y explained by x. That is, coefficient of determination is the ratio of explained variance to the total variance. For example, r2 = 0.879 means that 87.9% of the total variances in y are explained by x. When r2 = 1, it means that all the points on the scatter diagram fall on the regression line and the entire variations are explained by the straight line. On the other hand, if r2 = 0 it means that none of the points on scatter diagram falls on the regression line, meaning thereby that there is no linear relationship between the variables..

Example: Consider the following data:

X:	15	16	17	18	19	20
Y:	80	75	60	40	30	20

1. Fit both regression lines
2. Find the correlation coefficient
3. Verify the correlation coefficient is the geometric mean of the regression coefficients
4. Find the value of y when x = 17.5

Solution:

X	Y	XY	X^2	Y^2
15	80	1200	225	6400
16	75	1200	256	5625
17	60	1020	289	3600
18	40	720	324	1600
19	30	570	361	900
20	20	400	400	400
105	305	5110	1855	18525

$$\bar{x} = \frac{\Sigma x}{n} = \frac{105}{6} = 17.5, \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{305}{6} = 50.83$$

$$S_{xy} = \frac{1}{n}\Sigma x_i y_i - \bar{x}\,\bar{y} = \frac{5110}{6} - 17.5 \times 50.83 = -37.86$$

$$S_x^2 = \frac{1}{n}\Sigma x_i^2 - (\bar{x})^2 = \frac{1855}{6} - 17.5^2 = 2.92$$

$$S_y^2 = \frac{1}{n}\Sigma y_i^2 - (\bar{y})^2 = \frac{18525}{6} - 50.83^2 = 503.81$$

$$b_{yx} = \frac{S_{xy}}{S_x^2} = \frac{-37.86}{2.92} = -12.96 \text{ and } b_{xy} = \frac{S_{xy}}{S_y^2} = \frac{-37.86}{503.81} = -0.075$$

1. Regression line of y on x is $y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$

i.e., $y - 50.83 = -12.96(x - 17.5)$

$$\mathbf{y = -12.96\,x + 277.63}$$

Regression line of x on y is $x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$

i.e., $x - 17.5 = -0.075(y - 50.83)$

$$\mathbf{x = -0.075\,y + 21.31}$$

2. Correlation coefficient, $r = \frac{S_{xy}}{S_x S_y}$

$$= \frac{-37.86}{1.71 \times 22.45} = \mathbf{0.986}$$

3. $b_{yx} \times b_{xy} = -12.96 \times -0.075 = 0.972$

Then, $\sqrt{0.972} = \mathbf{0.986}$

So, $\mathbf{r = -0.986}$

4. To predict the value of y, use regression line of y on x.
When x= 17.5, $y = -12.96 \times 17.5 + 277.63 = \mathbf{50.83}$

Example 2

Calculate the correlation coefficient between X and Y from the following data:

$$\sum_{i=1}^{15}(X_i - \bar{X})^2 = 136 \quad \sum_{i=1}^{15}(Y_i - \bar{Y})^2 = 138 \quad \sum_{i=1}^{15}(X_i - \bar{X})(Y_i - \bar{Y}) = 122$$

Solution:

We have $$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}} = \frac{122}{\sqrt{136}\sqrt{138}}$$ r(X,Y) = 0.89

Example 3

Some health researchers have reported an inverse relationship between central nervous system malformations and the hardness of the related water supplies. Suppose the data were collected on a sample of 9 geographic areas with the following results:

C.N.S Malformation Rate(per1000births)	9	8	5	1	4	2	3	6	7

Water hardness(ppm)	120	130	90	150	160	100	140	80	200

Calculate the Correlation Coefficient between the C.N.S. malformation rate and Water hardness.

Solution:

Let us denote the C.N.S. malformation rate by x and water hardness by y. The mean of the x series = 5 and the mean of the y series =130, hence we can use the formula.

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}}$$

Calculation of correlation coefficient

x	Y	(x –)= x – 5	(y –) = y – 130	$(x -)^2$	$(y -)^2$	(x –) (y –)
9	120	4	– 10	16	100	– 40
8	130	3	0	9	0	0
5	90	0	– 40	0	1600	0
1	150	– 4	20	16	400	– 80
4	160	– 1	30	1	900	– 30
2	100	– 3	– 30	9	900	90
3	140	– 2	10	4	100	– 20
6	80	1	– 50	1	2500	– 50
7	200	2	70	4	4900	140
				Σ(x –)2= 60	Σ(y –)2= 11400	Σ(x –) (y –)= 10

=

r(X,Y) = 0.012

Therefore, the correlation coefficient between the C.N.S. malformation rate and water hardness is 0.012.

Example 4

Find the product moment correlation for the following data

X	57	62	60	57	65	60	58	62	56
Y	71	70	66	70	69	67	69	63	70

Solution:

X	**Y**	**XY**	**X^2**	**Y^2**
57	71	4047	3249	5041
62	70	4340	3844	4900
60	66	3960	3600	4356
57	70	3990	3249	4900
65	69	4485	4225	4761
60	67	4020	3600	4489
58	69	4002	3364	4761
62	63	3906	3844	3969
56	70	3920	3136	4900
537	**615**	**36670**	**32111**	**42077**

Thus we have, n = 9, $\sum X = 537$, $\sum Y = 615$, $\sum XY = 36670$, $\sum X^2 = 32111$, $\sum Y^2 = 42077$

Spearman's Rank Correlation Coefficient

If X and Y are qualitative variables then Karl Pearson's coefficient of correlation will be meaningless. In this case, we use Spearman's rank correlation coefficient which is defined as follows:

Where d is the difference in ranks.

Problems:

1. The ranks of same 16 students in Mathematics and Physics are as follows. The numbers within brackets denote the ranks of the students in Mathematics and Physics. (1,1), (2,10), (3,3), (4,4), (5,5), (6,7), (7,2), (8,6), (9,8), (10,11) (11. 15), (12,9), (13,14), (14,12), (15,16), (16,13). Calculate the rank correlation coefficient for the proficiencies of this group in Mathematics and Physics.

Solution:

Ranks in Maths(X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Ranks in Physics(Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
d = X – Y	0	–8	0	0	0	–1	5	2	1	–1	–4	3	–1	2	–1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Spearman's Rank Correlation Coefficient is given by,

= 0.8

2. The coefficient of rank correlation between the marks in Statistics and Mathematics obtained by a certain group of students is 2/3 and the sum of the squares of the differences in ranks is 55. Find the number of students in the group.

Solution:

Spearman's rank correlation coefficient is given by

Here $\rho = 2/3$, $\sum d^2 = 55$, N = ?

Therefore

Solving the above equation we get N = 10.

Repeated Ranks:

If any two or more individuals are equal in the series then Spearman's formula for calculating the rank correlation coefficients breaks down. In this case, common ranks are given to the repeated ranks. This common rank is the average of the ranks which these items would have assumed if they are slightly different from each other and the next item will get the rank next the ranks already assumed. As a result of this, following adjustment is made in the formula: add the factor to $\sum d^2$ where m is the number of items an item is repeated. This correction factor is to be added for each repeated value.

3. Obtain the rank correlation coefficient for the following data:

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

Solution:

X	**Y**	**Rank X**	**Rank Y**	**D = X – Y**	**D^2**
68	62	4	5	– 1	1
64	58	6	7	– 1	1
75	68	2.5	3.5	– 1	1
50	45	9	10	– 1	1
64	81	6	1	5	25
80	60	1	6	– 5	25
75	68	2.5	3.5	– 1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
					72

In X series 75 is repeated twice which are in the positions 2nd and 3rd ranks.

Therefore a common rank 2.5 (which is the average of 2 and 3) is given for each 75.

The corresponding correction factor is.

Also in the X series 64 is repeated thrice which are in the position 5^{th}, 6^{th} and 7^{th} ranks. Therefore, a common rank 6(which is the average of 5, 6 and 7) is given for each 64.

The corresponding correction factor is.

Similarly, in the Y series, 68 is repeated twice which are in the positions 3^{rd} and 4^{th} ranks. Therefore, a common rank (which is the average of 3 and 4) is given for each 68.

The corresponding correction factor is.

Rank correlation coefficient is

= 0.5454.

Example: Calculate the rank correlation coefficient from the sales and expenses of 10 firms are below:

Sales(X):	50	50	55	60	65	65	65	60	60	50
Expenses(Y):	11	13	14	16	16	15	15	14	13	13

Solution:

x	R_1	y	R_2	$d = R_1 - R_2$	d^2
50	9	11	10	-1	1
50	9	13	8	1	1
55	7	14	5.5	1.5	2.25
60	5	16	1.5	3.5	12.25
65	2	16	1.5	0.5	0.25
65	2	16	3.5	-1.5	2.25
65	2	15	3.5	-1.5	2.25
60	5	14	5.5	-0.5	0.25
60	5	13	8	-3	9
50	9	13	8	1	1
					31.5

Here there are 7 tied ranks, $m_1 = 3, m_2 = 3, m_3 = 3, m_4 = 2, m_5 = 2, m_6 = 2, m_7 = 3$

$$r = 1 - \frac{6[\Sigma d_i^2 + \Sigma \frac{1}{12}(m^3 - m)]}{n(n^2 - 1)}$$

$$= 1 - \frac{6[31.5 + \frac{1}{12}[(3^3 - 3) + (3^3 - 3) + (3^3 - 3) + (2^3 - 2) + (2^3 - 2) + (2^3 - 2) + (3^3 - 3)]]}{10(10^2 - 1)}$$

$$= 0.75$$

EXCERSIS

1. A company selling household appliances wants to determine if there is any relationship between advertising expenditures and sales. The following data was compiled for 6 major sales regions. The expenditure is in thousands of rupees and the sales are in millions of rupees.

Region :	1	2	3	4	5	6
Expenditure(X):	40	45	80	20	15	50
Sales (Y):	25	30	45	20	20	40

   a) Compute the line of regression to predict sales
   b) Compute the expected sales for a region where Rs.72000 is being spent on advertising
2. The following data represents the scores in the final exam., of 10 students, in the subjects of Economics and Finance.

Economics:	61	78	77	97	65	95	30	74	55
Finance:	84	70	93	93	77	99	43	80	67

   a) Compute the correlation coefficient?
3. Calculate the rank correlation coefficient from the sales and expenses of 9

Regression Analysis

Regression analysis helps us to estimate or predict the value of one variable from the given value of another. The known variable(or variables) is called independent variable(s). The variable we are trying to predict is the dependent variable.

Regression equations

Prediction or estimation of most likely values of one variable for specified values of the other is done by using suitable equations involving the two variables. Such equations are known as Regression Equations

Regression equation of y on x:

$y - = b_{yx} (x -)$ where y is the dependent variable and x is the independent variable and b_{yx} is given by

or

Regression equation of x on y:

$x - = b_{xy} (y -)$ where y is the dependent variable and x is the independent variable and b_{yx} is given by

or

b_{yx} and b_{xy} are called as regression coefficients of y on x and x on y respectively.

Relation between correlation and regression coefficients:

and

. = = r^2

Hence

Note: In the above expression the components inside the square root is valid only when b_{yx} and b_{xy} have the same sign. Therefore the regression coefficients will have the same sign.

Example 1

In trying to evaluate the effectiveness of its advertising campaign a company compiled the following information. Calculate the regression line of sales on advertising.

Year	1980	1981	1982	1983	1984	1985	1986	1987
Advertisement in 1000 rupees	12	15	15	23	24	38	42	48
Sales in lakes of rupees	5	5.6	5.8	7.0	7.2	88	9.2	9.5

Solution:

Let x be advertising amount and y be the sales amount.

Here, n = 8, ,

We know that, Regression equation of y on x is given by

$y - = b_{yx} (x -)$

Where

X	**Y**	**X^2**	**XY**

12	5	144	60
15	5.6	225	84
15	5.8	225	87
23	7.0	529	161
24	7.2	576	172.8
38	8.8	1444	334.4
42	9.2	1764	386.4
48	9.5	2304	456
217	**58.1**	**7211**	**1741.6**

Therefore

Substituting this value in the y on x equation, we get,

$$y - 7.26 = 0.125(x - 27.1)$$

Therefore the required equation of Sales on Advertisement is $y = 3.87 + 0.125\,x$

Example 2

In a study of the effect of a dietary component on plasma lipid composition, the following ratios were obtained on a sample of experimental anumals

Measure of dietary component (X)	1	5	3	2	1	1	7	3
Measure of plasma lipid level (Y)	6	1	0	0	1	2	1	5

(i) Obtain the two regression lines and hence predict the ratio of plasma lipid level with 4 dietary components.

(ii) Find the correlation coefficient between X and Y

Solution:

(i)

X	**Y**	**XY**	**X^2**	**Y^2**

1	6	6	1	36
5	1	5	25	1
3	0	0	9	0
2	0	0	4	0
1	1	1	1	1
1	2	2	1	4
7	1	7	49	1
3	5	15	9	25
23	**16**	**36**	**99**	**68**

Here n = 8 = 2.875 = 2

The Regression equation of y on x is given by $y - = b_{yx} (x -)$

Where

Hence the regression equation of y on x is

$y - 2 = - 0.304(x - 2.875)$

(i.e.) $y = 2.874 - 0.304 x$

When x = 4 (measure of dietary component) the plasmid lipid level is

$y = 2.874 - 0.304 (4)$

$y = 1.658$

The Regression equation of x on y is given by $x - = b_{xy} (y -)$

Where

Hence the regression equation of x on y is

$x - 2.875= - 0.278(y - 2)$

(i.e.) $x = 3.431 - 0.278 y$

(ii) The correlation coefficient between x and y is given by

RECALL

In this Lesson the concept of correlation and regression are discussed. The correlation is the association between two variables. A scatter plot of the variables may suggest that the two variables are related but the value of the Pearson's correlation coefficient r quantifies this association. The correlation coefficient r may assume values from –1 and + 1. The sign indicates whether the association is direct (+ve) or inverse (-ve). A numerical value of 1 indicates perfect association while a value of zero indicates no association. Regression is a device for establishing relationships between variables from the given data. The discovered relationship can be used for predictive purposes. Some simple examples are shown to understand the concepts.

References

1. P.R. Vital – Business Mathematics and Statistics.

2. Gupta S.P. – Statistical Methods.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

SMTA1104 – Business Statistics

UNIT – IV – Probability – SMTA1104

Course Material
Business Statistics

Unit - IV
Probability

Introduction

If an experiment is repeated under essential homogeneous and similar conditions we generally come across two types of situations:

(i) The result or what is usually known as the 'outcome' is unique or certain.

(ii) The result is not unique but may be one of the several possible outcomes.

The phenomena covered by (i) are known as deterministic. For example, for a perfect gas, PV = constant.

The phenomena covered by (ii) are known as probabilistic. For example, in tossing a coin we are not sure if a head or tail will be obtained.

In the study of statistics we are concerned basically with the presentation and interpretation of chance outcomes that occur in a planned study or scientific investigation.

Definition of various terms

Trial and event: Consider an experiment which, though repeated under essentially identical conditions, does not give unique results but may result in any one of the several possible outcomes. The experiment is known as a trial and outcomes are known as events or cases. For example, throwing of a die is a trial and getting 1(or 2 or … 6) is an event.

Exhaustive events: The total number of possible outcomes in any trial is known as exhaustive events or exhaustive cases. For example, in tossing of a coin there are two exhaustive case, viz.: Head and Tail(the possibility of the coin standing on an edge being ignored)

Favourable events or cases: The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event. For example, in throwing of two dice, the number of cases favourable to getting the sum 3 is: (1,2) and (2,1)

Mutually exclusive events: Events are said to be mutually exclusive or incompatible if the happening of any one of them precludes the happening of all the others, that is if no two or more of them can happen simultaneously in the same trial. For example, in tossing a coin the events head and tail are mutually exclusive.

Equally likely events: Outcomes of a trial are said to be equally likely, if taking into consideration all the relevant evidences, there is no reason to expect one in preference to the others. For example, in throwing an unbiased die, all the six faces are equally likely to come.

Sample Space: Consider an experiment whose outcome is not predictable with certainty. However, although the outcome of the experiment will not be known in advance, let us suppose that the set of all possible outcomes is known. This set of all possible outcomes of an experiment is known as the **sample space** of the experiment and is denoted by S.

Some examples follow.

1. If the outcome of an experiment consists in the determination of the sex of a newborn child, then

   $S = \{ g,b\}$

   where the outcome g means that the child is a girl and b that it is a boy.

2. If the experiment consists of flipping two coins, then the sample space consists of the following four points:

   $S = \{(H,H), (H,T), (T,H), (T,T)\}$

   The outcome will be (H,H) if both coins are heads, (H,T) if the first coin is heads and the second tails, (T,H) if the first is tails and the second heads, and (T,T) if both coins are tails.

3. If the experiment consists of tossing two dice, then the sample space consists if the 36 points

   $S = \{ (i,j):i,j = 1, 2, 3, 4, 5,$

   $= \{ (1,1)$------$(1,6)$-----$(6,1)$-----$(6,6) \}$

   where the outcome (i,j) is said to occur if i appears on the leftmost die and j on the other die.

3.2. Definitions of Probability

1. Mathematical or Classical or a priori probability:

If a trial results in n exhaustive, mutually exclusive and equally likely cases and m of them are favourable to the happening of an event E, then the probability 'p' of happening of E is given by,

$$p = P(E) = \frac{\text{Favourable number of cases}}{\text{Exhaustive number of cases}} = \frac{m}{n}$$

2. Statistical or empirical probability:

If a trial is repeated a number of times under essentially homogenous and identical conditions, then the limiting value of the number of times the event happens to the number of trials, as the number of trials become indefinitely large is called the probability of happening of the event. Symbolically, if in n trials an event E happens m times, then the probability 'p' of the happening of E is given by,

$$P = P(E) = \lim_{n\to\infty} \frac{m}{n}$$

3. Axiomatic Definition:

Consider an experiment whose sample space is S. For each event E of the sample space S, we assume that a number P(E) is defined and satisfies the following three axioms.

Axiom 1: $0 \le P(E) \le 1$

Axiom 2: $P(S) = 1$

Axiom 3: For any sequence of mutually exclusive events, $E_1, E_2, \ldots$ (that is, events for which $E_i E_j = \Phi$, when $i \neq j$),

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Some Important Formulas

1. If A and B are any two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This rule is known as additive rule on probability.

For three events A, B and C, we have,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

2. If A and B are mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

In general, if $A_1, A_2, \ldots, A_n$ are mutually exclusive, then

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

3. If A and A^c are complementary events, then

$$P(A) + P(A^c) = 1$$

4. $P(S) = 1$

5. $P(\Phi) = 0$

6. If A and B are any two events, then

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

7. If A and B are independent events, then
$$P(A \cap B) = P(A) \times P(B)$$

Glossary of Probability terms:

Statement	Meaning in terms of Set theory
1. At least one of the events A or B occurs	$\omega \in A \cup B$
2. Both the events A and B occur	$\omega \in A \cap B$
3. Neither A nor B occurs	$\omega \in \overline{A} \cap \overline{B}$
4. Event A occurs and B does not occur	$\omega \in A \cap \overline{B}$
5. Exactly one of the events A or B occurs	$\omega \in A \Delta B$
6. If event A occurs, so does B	$\mathbf{A \subset B}$
7. Events A and B are mutually exclusive	$A \cap B = \Phi$
8. Complementary event of A	$\overline{A}$
9. Sample space	**Universal set S**

Example 1: Find the probability of getting a head in tossing a coin.
Solution: When a coin is tossed, we have the sample space {Head, Tail}
Therefore, the total number of possible outcomes is 2
The favourable number of outcomes is 1, that is the head.
$\therefore$ The required probability is ½.

Example 2: Find the probability of getting two tails in two tosses of a coin.
Solution: When two coins are tossed, we have the sample space {HH, HT, TH, TT}
Where H represents the outcome Head and T represents the outcome Tail.
The total number of possible outcomes is 4.
The favourable number of outcomes is 1, that is TT
$\therefore$ The required probability is ¼.

Example 3: Find the probability of getting an even number when a die is thrown
Solution: When a die is thrown the sample space is {1, 2, 3, 4, 5, 6}
The total number of possible outcomes is 6
The favourable number of outcomes is 3, that is 2, 4 and 6
$\therefore$ The required probability is= $\frac{3}{6}$ = ½.

Example 4: What is the chance that a leap year selected at random will contain 53 Sundays?
Solution: In a leap year(which consists of 366 days) there are 52 complete weeks and 2 days over. The following are the possible combinations for these two over days:

(i) Sunday and Monday (ii)Monday and Tuesday (iii)Tuesday and Wednesday (iv)Wednesday and Thursday (v)Thursday and Friday (vi)Friday and Saturday (vii)Saturday and Sunday.

In order that a leap year selected at random should contain 53 Sundays, one of the two over days must be Sunday. Since out of the above 7 possibilities, 2 viz. (i) and (ii)are favourable to this event,

$$\text{Required probability} = \frac{2}{7}$$

Example 5: If two dice are rolled, what is the probability that the sum of the upturned faces will equal 7?

Solution: We shall solve this problem under the assumption that all of the 36 possible outcomes are equally likely. Since there are 6 possible outcomes – namely (1,6), (2,5), (3,4), (4,3), (5,2,), (6,1) – that result in the sum of the dice being equal to 7, the desired probability is $\frac{6}{36} = \frac{1}{6}$.

Example 6: A bag contains 3 Red, 6 White and 7 Blue balls. What is the probability that two balls drawn are white and blue?

Solution: Total number of balls = 3 + 6 + 7 = 16.

Out of 16 balls, 2 can be drawn in $16C_2$ ways.

Therefore exhaustive number of cases is 120.

Out of 6 white balls 1 ball can be drawn in $6C_1$ ways and out of 7 blue balls 1 ball can be drawn in $7C_1$ ways. Since each of the former cases can be associated with each of the latter cases, total number of favourable cases is $6C_1 \times 7C_1 = 6 \times 7 = 42$.

$\therefore$ The required probability is $= \frac{42}{120} = \frac{7}{20}$

Example 7: A lot consists of 10 good articles, 4 with minor defects and 2 with major defects. Two articles are chosen from the lot at random (without replacement). Find the probability that (i) both are good, (ii) both have major defects, (iii) at least 1 is good, (iv) at most 1 is good, (v)exactly 1 is good, (vi) neither has major defects and (vii) neither is good.

Solution: Although the articles may be drawn one after the other, we can consider that both articles are drawn simultaneously, as they are drawn without replacement.

(i) $$P(\textit{both}\ \text{are good}) = \frac{\text{No. of ways drawing 2 good articles}}{\text{Total no. of ways of drawing 2 articles}}$$

$$= \frac{10C_2}{16C_2} = \frac{3}{8}$$

(ii)

$$P(\text{both have major defects}) = \frac{\text{No. of ways of drawing 2 articles with major defects}}{\text{Total no. of ways}}$$

$$= \frac{2C_2}{16C_2} = \frac{1}{120}$$

(iii) P(at least 1 is good) = P(exactly 1 is good or both are good)

=P(exactly 1 is good and 1 is bad or both are good)

$$= \frac{10C_1 x 6C_1 + 10C_2}{16C_2} = \frac{7}{8}$$

(iv) P(atmost 1 is good) =P(none is good or 1 is good and 1is bad)

$$= \frac{10C_0 x 6C_2 + 10C_1 x 6C_1}{16C_2} = \frac{5}{8}$$

(v) P(exactly 1is good) =P(1 is good and 1 is bad)

$$= \frac{10C_1 x 6C_1}{16C_2} = \frac{1}{2}$$

(vi) P(neither has major defects) = P(both are non-major defective articles)

$$= \frac{14C_2}{16C_2} = \frac{91}{120}$$

(vii) P(neither is good) = P(both are defective)

$$= \frac{6C_2}{16C_2} = \frac{1}{8}$$

Example 8: From 6 positive and 8 negative numbers, 4 numbers are chosen at random (without replacement) and multiplied. What is the probability that the product is positive?

Solution: If the product is to be positive, all the 4 numbers must be positive or all the 4 must be negative or 2 of them must be positive and the other 2 must be negative.

No. of ways of choosing 4 positive numbers= $6C_4 = 15$.

No. of ways of choosing 4 negative numbers= $8C_4 = 70$.

No.of ways of choosing 2 positive and 2 negative numbers

$= 6C_2 x 8C_2 = 420$.

Total no. of ways of choosing 4 numbers from all the 14 numbers

$= 14C_4 = 1001$.

P(the product is positive)

$$= \frac{No. \text{ of ways by which t he product is positive}}{\text{Total no. of ways}}$$

$$= \frac{15 + 70 + 420}{1001} = \frac{505}{1001}$$

Example 9: If 3 balls are "randomly drawn" from a bowl containing 6 white and 5 black balls, what is the probability that one of the drawn balls is white and the other two black?

Solution: If we regard the order in which the balls are selected as being relevant, then the sample space consists of $11 \cdot 10 \cdot 9 = 990$ outcomes. Furthermore, there are $6 \cdot 5 \cdot 4 = 120$ outcomes in which the first ball selected is white and the other two black; $5 \cdot 6 \cdot 4 = 120$ outcomes in which the first is black, the second white and the third black; and $5 \cdot 4 \cdot 6 = 120$ in which the first two are black and the third white. Hence, assuming that "randomly drawn" means that each outcome in the sample space is equally likely to occur, we see that the desired probability is $\frac{120 + 120 + 120}{990} = \frac{4}{11}$

Example 10: In a large genetics study utilizing guinea pigs, *Cavia sp.,* 30% of the offspring produced had white fur and 40% had pink eyes. Two-thirds of the guinea pigs with white fur had pink eyes. What is the probability of a randomly selected offspring having both white fur and pink eyes?

Solution: P(W) = 0.30, P(Pi) = 0.40, and P(Pi W) = 0.67. Utilizing Formula 2.9,

$$P(Pi \cap W) = P(Pi \mid W). P(W) = 0.67. 0.30 = 0.20.$$

Twenty percent of all offspring are expected to have both white fur and pink eyes.

Example 11: Consider three gene loci in tomato, the first locus affects fruit shape with the oo genopyte causing oblate or flattened fruit and OO or Oo normal round fruit. The second locus affects fruit color with yy having yellow fruit and YY or Yy red fruit. The final locus affects leaf shape with pp having potato or smooth leaves and PP or Pp having the more typical cut leaves. Each of these loci is located on a different pair of chromosomes and, therefore, acts independently of the other loci. In the following cross OoYyPp × OoYypp, what is the probability that an offspring will have the dominant phenotype for each trait? What is the probability that it will be heterozygous for all three genes? What is the probability that it will have round, yellow fruit and potato leaves?

Solution: Genotypic array:

$$(\frac{1}{4}\text{OO} + \frac{2}{4}\text{Oo} + \frac{1}{4}\text{oo})\,(\frac{1}{4}\text{YY} + \frac{2}{4}\text{Yy} + \frac{1}{4}\text{yy})\,(\frac{1}{2}\text{pp})$$

Phenotypic array:

$$(\frac{3}{4}\text{O-} + \frac{1}{4}\text{oo})\,(\frac{3}{4}\text{Y-} + \frac{1}{4}\text{yy})\,(\frac{1}{2}\text{P} + \frac{1}{2}\text{pp})$$

The probabiltity of dominant phenotype for each trait from the phenotypic array above is

$$P(\text{O-Y-P-}) = P(\text{O-}) \times P(\text{Y-}) \times P(\text{P-}) = \frac{3}{4} \times \frac{3}{4} \times \frac{1}{2} = \frac{9}{32}.$$

The probability of heterozygous for all three genes from the genotypic array above is

$$P(OoYyPp) = P(Oo) \times P(Yy) \times P(Pp) = \frac{2}{4} \times \frac{2}{4} \times \frac{1}{2} = \frac{4}{32} = \frac{1}{8}.$$

The probability of a round, yellow-fruited plant with potato leaves from the phenotypic array above is

$$P(O\text{-}yypp) = P(O\text{-}) \times P(yy) \times P(pp) = \frac{3}{4} \times \frac{1}{4} \times \frac{1}{2} = \frac{3}{32}.$$

Each answer applies the probability rules for independent events to the separate gene loci.

Example 12: (a) Two cards are drawn at random from a well shuffled pack of 52 playing cards. Find the chance of drawing two aces.

(b) From a pack of 52 cards, three are drawn at random. Find the chance that they are a king, a queen and a knave.

(c) Four cards are drawn from a pack of cards. Find the probability that (i) all are diamond (ii) there is one card of each suit (iii) there are two spades and two hearts.

Solution: (a) From a pack of 52 cards 2 can be drawn in $52C_2$ ways, all being equally likely. $\therefore$ Exhaustive number of cases is $52C_2$.

In a pack there are 4 aces and therefore 2 aces can be drawn in $4C_2$ ways.

$\therefore$ Required probability = $\frac{4C_2}{52C_2} = \frac{1}{221}$

(b) Exhaustive number of cases = $52C_3$

A pack of cards contains 4 kings, 4 queens and 4 knaves. A king, a queen and a knave can each be drawn in $4C_1$ ways and since each way of drawing a king can be associated with each of the ways of drawing a queen and a knave, the total number of favrourable cases = $4C_1 \times 4C_1 \times 4C_1$.

$\therefore$ Required probability = $\frac{4C_1 \times 4C_1 \times 4C_1}{52C_3} = \frac{16}{5525}$

(c) Exhaustive number of cases $52C_4$

(i) Required probability = $\frac{13C_4}{52C_4}$

(ii) Required probability = $\frac{13C_1 \times 13C_1 \times 13C_1 \times 13C_1}{52C_4}$

(iv) Required probability = $\frac{13C_2 \times 13C_2}{52C_4}$

Example 13: What is the probability of getting 9 cards of the same suit in one hand at a game of bridge?

Solution: One hand in a game of bridge consists of 13 cards.

$\therefore$ Exhaustive number of cases $52C_{13}$

Number of ways in which, in one hand, a particular player gets 9 cards of one suit are $13C_9$ and the number of ways in which the remaining 4 cards are of some other suit are $39C_4$. Since there are 4 suits in a pack of cards, total number of favourable cases is $4 \times 13C_9 \times 39C_4$.

$\therefore$ Required probability = $\frac{4 \times 13C_9 \times 39C_4}{52C_{13}}$

Example 14: A committee of 4 people is to be appointed from 3 officers of the production department, 4 officers of the purchase department, two officers of the sales department and 1 chartered accountant. Find the probability of forming the committee in the following manner:

(i) There must be one from each category
(ii) It should have at least one from the purchase department
(iii) The chartered accountant must be in the committee.

Solution: There are 3 + 4 + 2 + 1 = 10 persons in all and a committee of 4 people can be formed out of them in $10C_4$ ways. Hence exhaustive number of cases is $10C_4$ = 210

(i) Favourable number of cases for the committee to consist of 4 members, one from each category is $4C_1 \times 3C_1 \times 2C_1 \times 1 = 24$

$\therefore$ Required probability = $\frac{24}{120}$

(ii) P(Committee has at least one purchase officer) = 1 – P(Committee has no purchase Officer)

In order that the committee has no purchase officer, all the four members are to be selected amongst officers of production department, sales department and chartered accountant, that is out of 3 + 2 + 1 = 6 members and this can be done in $5C_4$ = 15 ways. Hence,

P(Committee has no purchase officer) = $\frac{15}{210} = \frac{1}{14}$

$\therefore$ P(Committee has at least one purchase officer) = $1 - \frac{1}{14} = \frac{13}{14}$

(iii) Favourable number of cases that the committee consists of a chartered accountant as a member and three others are:

$1 \times {}^{9}C_{3} = 84$ ways.

Since a chartered accountant can be selected out of one chartered accountant in only 1 way and the remaining 3 members can be selected out of the remaining 10 – 1 persons in ${}^{9}C_{3}$ ways. Hence the required probability = $\frac{84}{210} = \frac{2}{5}$.

Example 15: A box contains 6 red, 4 white and 5 black balls. A persons draws 4 balls from the box at random. Find the probability that among the balls drawn there is at least one ball of each colour.

Solution: The required event E that in a draw of 4 balls from the box at random there is at least one ball of each colour can materialize in the following mutually disjoint ways:

(i) 1 Red, 1 White and 2 Black balls
(ii) 2 Red, 1 White and 1 Black balls
(iii) 1 Red, 2 White and 1 Black balls

Hence by addition rule of probability, the required probability is given by,

$$P(E) = P(i) + P(ii) + P(iii)$$

$$= \frac{6C_1 \times 4C_1 \times 5C_2}{15C_4} + \frac{6C_2 \times 4C_1 \times 5C_1}{15C_4} + \frac{6C_1 \times 4C_2 \times 5C_1}{15C_4}$$

$$= 0.5275$$

Example 16: A problem in Statistics is given to the three students A, B and C whose chances of solving it are 1/2, 3/4 and 1/4 respectively. What is the probability that the problem will be solved if all of them try independently?

Solution: Let A, B and C denote the events that the problem is solved by the students A, B and C respectively. Then

$P(A) = 1/2$ $P(B) = 3/4$ $P(C) = 1/4$

$P(\overline{A}) = 1 - 1/2 = 1/2$ $P(\overline{B}) = 1 - 3/4 = 1/4$ $P(\overline{C}) = 1 - 1/4 = 3/4$

P(Problem solved) = P(At least one of them solves the problem)

$$= 1 - P(\text{None of them solve the problem})$$

$$= 1 - P(\overline{A \cup B \cup C})$$

$$= 1 - P(\overline{A} \cap \overline{B} \cap \overline{C})$$

$$= 1 - P(\overline{A})\, P(\overline{B})\, P(\overline{C})$$

$$= 1 - \frac{1}{2} \times \frac{1}{4} \times \frac{3}{4}$$

$$= \frac{29}{32}$$

Example 17: Three groups of children contain respectively 3 girls and 1 boy, 2 girls and 2 boys and 1 girl and 3 boys. One child is selected at random from each group. Find the probability that the three selected consist of 1 girl and 2 boys.

Solution: The required event of getting 1 girl and 2 boys among the three selected children can materialize in the following three mutually exclusive cases:

Group No. →	**I**	**II**	**III**
(i)	**Girl**	**Boy**	**Boy**
(ii)	**Boy**	**Girl**	**Boy**
(iii)	**Boy**	**Boy**	**Girl**

By addition rule of probability,

Required probability = P(i) + P(ii) + P(iii)

Since the probability of selecting a girl from the first group is 3/4, of selecting a boy from the second is 2/4, and of selecting a boy from the third group is ¾, and since these three events of selecting children from the three groups are independent of each other, we have,

$$P(i) = \frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{9}{32}$$

$$P(ii) = \frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{3}{32}$$

$$P(iii) = \frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{1}{32}$$

Hence the required probability $= \frac{9}{32} + \frac{3}{32} + \frac{1}{32} = \frac{13}{32}$

Conditional Probability and Baye's Theorem

Conditional Probability and Multiplication Law

For two events A and B

$P(A \cap B) = P(A) \cdot P(B/A), P(A) > 0$

$= P(B) \cdot P(A/B), P(B) > 0$

where P(B/A) represents the conditional probability of occurrence of B when the event A has already happened and P(A/B) is the conditional probability of occurrence of A when the event B has already happened.

Theorem of Total Probability:

If $B_1, B_2, \ldots, B_n$ be a set of exhaustive and mutually exclusive events, and A is another event associated with (or caused by) B_i, then

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A/B_i)$$

Example 18 : A box contains 4 bad and 6 good tubes. Two are drawn out from the box at a time. One of them is tested and found to be good. What is the probability that the other one is also good?

Solution: Let A = one of the tubes drawn is good and B = the other tube is good.

$P(A \cap B)$ = P(both tubes drawn are good)

$$= \frac{6C_2}{10C_2} = \frac{1}{3}$$

Knowing that one tube is good, the conditional probability that the other tube is also good is required, i.e., P(B/A) is required.

By definition,

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/3}{6/10} = \frac{5}{9}$$

Example 19: A bolt is manufactured by 3 machines A, B and C. A turns out twice as many items as B, and machines B and C produce equal number of items. 2% of bolts produced by A and B are defective and 4% of bolts produced by C are defective. All bolts are put into 1 stock pile and chosen from this pile. What is the probability that it is defective?

Solution: Let A = the event in which the item has been produced by machine A, and so on.

Let D = the event of the item being defective.

$$P(A) = \frac{1}{2}, \quad P(B) = P(C) = \frac{1}{4}$$

P(D/A) = P(an item is defective, given that A has produced it)

$$= \frac{2}{100} = P(D/B)$$

$$P(D/C) = \frac{4}{100}$$

By theorem of total probability,

$$P(D) = P(A) \times P(D/A) + P(B) \times P(D/B) + P(C) \times P(D/c)$$

$$= \frac{1}{2} \times \frac{2}{100} + \frac{1}{4} \times \frac{2}{100} + \frac{1}{4} \times \frac{4}{100}$$

$$= \frac{1}{40}$$

Example 20: In a coin tossing experiment, if the coin shows head, one die is thrown and the result is recorded. But if the coin shows tail, 2 dice are thrown and their sum is recorded. What is the probability that the recorded number will be 2?

Solution: When a single die is thrown, P(2) = 1/6

When 2 dice are thrown, the sum will be 2 only if each dice shows 1.

$\therefore$ P(getting 2 as sum with 2 dice) = $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ (since independence)

By theorem of total probability,

$$P(2) = P(H) \times P(2/H) + P(T) \times P(2/T)$$
$$= \frac{1}{2} \times \frac{1}{6} + \frac{1}{2} \times \frac{1}{36} = \frac{7}{72}$$

Example 21: An urn contains 10 white and 3 black balls. Another urn contains 3 white and 5 black balls. Two balls are drawn at random from the first urn and place in the second urn and then one ball is taken at random from the latter. What is the probability that it is a white ball?

Solution: The two balls transferred may be both white or both black or one white and one black.

Let B_1 = event of drawing 2 white balls from the first urn, B_2 = event of drawing 2 black balls from it and B_3 = event of drawing one white and one black ball from it.

Clearly B_1, B_2 and B_3 are exhaustive and mutually exclusive events.

Let A = event of drawing a white ball from the second urn after transfer.

$$P(B_1) = \frac{10C_2}{13C_2} = \frac{15}{26}$$

$$P(B_2) = \frac{3C_2}{13C_2} = \frac{1}{26}$$

$$P(B_3) = \frac{10 \times 3}{13C_2} = \frac{10}{26}$$

$P(A/B_1)$ = P(drawing a white ball / 2 white balls have been transferred)

= P(drawing a white ball / urn II contains 5 white and 5 black balls)

$$= \frac{5}{10}$$

Similarly, $P(A/B_2) = \frac{3}{10}$ and $P(A/B_3) = \frac{4}{10}$

By theorem of total probability,

$$P(A) = P(B_1) \times P(A/B_1) + P(B_2) \times P(A/B_2) + P(B_3) \times P(A/B_3)$$
$$= \frac{15}{26} \times \frac{5}{10} + \frac{1}{26} \times \frac{3}{10} + \frac{10}{26} \times \frac{4}{10} = \frac{59}{130}$$

Example 22: In 1989 there were three candidates for the position of principal – Mr.Chatterji, Mr. Ayangar and Mr. Singh – whose chances of getting the appointment are in the proportion 4:2:3 respectively. The probability that Mr. Chatterji if selected would introduce co-education in the college is 0.3. The probabilities of Mr. Ayangar

and Mr.Singh doing the same are respectively 0.5 and 0.8. What is the proabability that there will be co-education in the college?

Solution: Let the events and probabilities be defined as follows:

A: Introduction of co-education

E_1: Mr.Chatterji is selected as principal

E_2: Mr.Ayangar is selected as principal

E_3: Mr.Singh is selected as principal

Then,

$P(E_1) = \frac{4}{9}$ $P(E_2) = \frac{2}{9}$ $P(E_3) = \frac{3}{9}$

$P(A/E_1) = 0.3$ $P(A/E_2) = 0.5$ $P(A/E_3) = 0.8$

$$P(A) = P\left[(A \cap E_1) \cup (A \cap E_2) \cup (A \cap E_3)\right]$$
$$= P\left[(A \cap E_1) + (A \cap E_2) + (A \cap E_3)\right]$$
$$= P(E_1)\,P(A/E_1) + P(E_2)\,P(A/E_2) + P(E_3)\,P(A/E_3)$$
$$= \frac{4}{9} \times \frac{3}{10} + \frac{2}{9} \times \frac{5}{10} + \frac{3}{9} \times \frac{8}{10} = \frac{23}{45}$$

3.3.4. Baye's theorem

If $E_1, E_2, \ldots, E_n$ are mutually disjoint events with $P(E_i) \neq 0$, $(i = 1,2, \ldots, n)$ then for any arbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$, we have,

$$P(E_i/A) = \frac{P(E_i)P(A/E_i)}{\sum_{i=1}^{n} P(E_i)P(A/E_i)}, \quad i = 1, 2, \ldots, n$$

3.3.5. Solved Examples

Example 23. A bag contains 5 balls and it is not known how many of them are white. Two balls are drawn at random from the bag and they are noted to be white. What is the chance that all the balls in the bag are white?

Solution: Since 2 white balls have been drawn out, the bag must have contained 2, 3, 4 or 5 white balls.

Let B_1 = Event of the bag containing 2 white balls, B_2 = Events of the bag containing 3 white balls, B_3 = Event of the bag containing 4 white balls and B_4 = Event of the bag containing 5 white balls.

Let A = Event of drawing 2 white balls.

$P(A/B_1) = \frac{2C_2}{5C_2} = \frac{1}{10}$ $P(A/B_2) = \frac{3C_2}{5C_2} = \frac{3}{10}$

$P(A/B_3) = \frac{4C_2}{5C_2} = \frac{4}{10}$ $P(A/B_4) = \frac{5C_2}{5C_2} = 1$

Since the number of white balls in the bag is not known, B_i's are equally likely.

$$P(B_1) = P(B_2) = P(B_3) = P(B_4) = \frac{1}{4}$$

By Baye's theorem,

$$P(B_4/A) = \frac{P(B_4) \times P(A/B_4)}{\sum_{i=1}^{4} P(B_i) \times P(A/B_i)}$$

$$= \frac{\frac{1}{4} \times 1}{\frac{1}{4} \times \left(\frac{1}{10} + \frac{3}{10} + \frac{3}{5} + 1 \right)} = \frac{1}{2}$$

Example 24: There are 3 true coins and 1 false coin with 'head' on both sides. A coin is chosen at random and tossed 4 times. If 'head' occurs all the 4 times, what is the probability that the false coin has beeb chosen and used?

Solution:

$$P(T) = P(\text{the coin is a true coin}) = \frac{3}{4}$$

$$P(F) = P(\text{the coin is a false coin}) = \frac{1}{4}$$

Let A = Event of getting all heads in 4 tosses

Then $P(A/T) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$ and $P(A/F) = 1$

By Baye's theorem

$$P(F/A) = \frac{P(F) \times P(A/F)}{P(F) \times P(A/F) + P(T) \times P(A/T)}$$

$$= \frac{\frac{1}{4} \times 1}{\frac{1}{4} \times 1 + \frac{3}{4} \times \frac{1}{16}} = \frac{16}{19}$$

Example 25: The contents of urns I, Ii and III are as follows:

1 white, 2 black and 3 red balls
2 white, 1 black and 1 red balls
4 white, 5 black and 3 red balls

One urn is chosen at random and two balls are drawn. They happen to be white and red. What is the probability that they come from urns I, II or III?

Solution: Let E_1, E_2 and E_3 denote the events that the urn I, II and III is chosen, respectively, and let A be the event that the two balls taken from the selected urn are white and red. Then

$$P(E_1) = P(E_2) = P(E_3) = \frac{1}{3}$$

$$P(A/E_1) = \frac{1 \times 3}{6C_2} = \frac{1}{5}$$

$$P(A/E_2) = \frac{2 \times 1}{4C_2} = \frac{1}{3}$$

$$P(A/E_3) = \frac{4 \times 3}{12C_2} = \frac{2}{11}$$

Hence $P(E_2/A) = \dfrac{P(E_2)P(A/E_2)}{\sum_{i=1}^{3} P(E_i)P(A/E_i)}$

$$= \frac{\frac{1}{3} \times \frac{1}{3}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{11}} = \frac{55}{118}$$

Similarly, $P(E_3/A) = \dfrac{\frac{1}{3} \times \frac{2}{11}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{11}} = \frac{30}{118}$

Therefore $P(E_1/A) = 1 - \frac{55}{118} - \frac{30}{118} = \frac{33}{118}$

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

SMTA1104 – Business Statistics

UNIT – V – Time Series – SMTA1104

Course Material (B.Com.)
Subject Name - Business Statistics

UNIT-5

TIME SERIES ANALYSIS

INTRODUCTION:

A **time series** is a set of observations taken at specified times, usually at equal intervals. In other words, a *series of observations recorded over time is known as a* **time series**. Examples of time series are the data regarding population of a country recorded at the ten-yearly censuses, annual production of a crop, say, wheat over a number of years, the wholesale price index over a number of months, the daily closing price of a share on the stock exchange, the hourly temperature recorded by weather bureau of a city, the total monthly sales receipts in business establishment, and so on. *In fact, data related with business and economic activities, in general, recorded over time give* rise to a **time series.**

One of the most important tasks before the planners and administrators in the field of economic and business activities is to make future estimates based on the past behaviour of a phenomenon under consideration. For example, trade cycles are important to economists and others in business and commerce. The behaviour of the cycles and their causes are of interest to them. Such studies are to be based on the analysis of time series data collected over time. *Thus,* **the analysis of time series** *plays an important role in empirical investigations of economic, commercial, social and even biological phenomena.*

Mathematically, **a time series** is defined by the fractional relationship

$$Y_t = f(t)$$

where Y_t is the value of the variable (or phenomenon) under consideration over time t. Thus, if the values of a variable at time points $t_1, t_2, ..., t_n$ are $Y_1, Y_2, ..., Y_N$ respectively, then the series

t	:	t_1	t_2	$t_3 ....., t_N$
Y_t	:	Y_1	Y_2	$Y_3 ....., Y_N$

constitute a time series.

COMPONENTS OF TIME SERIES:

Empirical studies of a number of time series have revealed the presence of certain **characteristic movements or fluctuations** in a time series. *These characteristic movements of a time series may be classified in four different categories called* **components of time series.** In a long time series, generally, we have the following **four components :**

1. Secular Trend or long–term movements
2. Seasonal variations
3. Cyclic variations
4. Random or Irregular movements

SECULAR TREND:

Secular trend *means the general long–term tendency of a series. In fact, secular trend is that characteristic of a time series which extends consistently throughout the entire period of time under consideration. It shows a long–term tendency of an activity to grow or to decline.* For example, a time series on population shows a tendency to increase; time series of sales of a product shows a tendency to increase, and so on. On the other hand, a downward tendency is observed in the time series on birth and death rates. The factors which remain more or less constant over a long period also produce a trend. *The term* '**long period of time**' *is a relative phenomenon and cannot be defined exactly.* For some cases, a period as small as a week may be fairly long while in other cases, a period as long as 2 years may not be assumed long. For example, an increase in agricultural production over a period of two years would not be termed as secular change, whereas if the count of bacterial population of culture every five minutes, for a week shows an increase, then we would consider it as a secular change.

SEASONAL VARIATION:

The component responsible for the regular rise and fall in the magnitude of the time series is called **seasonal variation.** *In other words* **seasonal movements** or **seasonal variations** *refer to identical, or almost identical, patterns which a time series appears to follow during corresponding months of successive years.* Such variations are due to recurring events which takes place annually, quarterly, monthly, weekly or even daily, depending on the type of data available. *But in no case this period is to exceed one year. In view of their regular nature, seasonal variations are precise and can be foreseen,* as for instance the prices of agricultural commodities fall every year during the harvesting period, the sale of umbrellas pick up very fast in a rainy season, the demand for electric fans goes up during summer. Seasonal variations in general refer to annual periodicity in business and economic activities. These are the effects of seasonal factors like climatic conditions, human habits, fashions, customs and conventions of the people in a particular society.

CYCLICAL VARIATION:

Cyclical movements or **variations** *refer to the long-term oscillations or swings about a trend line. These cycles may or may not be periodic, i.e., they may or may not follow exactly similar patterns after equal intervals of time. Such variations are of longer duration than a year and they do not show the type of regularity as observed in the case of seasonal variations.* An important example of cyclical variations are the so-called **business cycles** representing intervals of **prosperity, recession, depression** and **recovery**. Each phase changes gradually into the phase which follows it in the given order. In a business activity, these phases follow each other with steady regularity and the period from the peak of one boom to the peak of the next boom is called a **complete cycle.** The usual periods of a business cycle may be ranging between 5–11 years. Most of the economic and business series relating to income, investment, wages, production shows this tendency.The study of cyclical fluctuations is therefore very important for predicting the turning phases in a business activity which may greatly help in proper policy formation in the area.

IRREGULAR VARIATION:

Random or Irregular *movements refer to such variations in a time series which do not repeat in a definite pattern.* Irregular movements in a time series may be of two types :

(i) Random or chance variations

(ii) Episodic variations

Random or chance variations *in a real phenomenon are inevitable by nature. It does effect a series in a random way, and as such, the effect of chance or random variations on a series is small.*

On the other hand, **episodic variations** *in a time series arise due to specific events or episodes like epidemic,* fire, strike or natural calamities like flood, earthquake or late monsoon etc. In some cases, irregular variations may not have a significant importance while in others these may be so intense as to result in new cyclical variations.

MEASUREMENT OF TREND:

The main objective behind the study of the trend of a time series are :

1. to describe the long–term growing or declining trend in a phenomenon under study.
2. to eliminate the trend component in order to bring into focus the remaining components in the time series.

In order to meet these objectives, some statistical methods of **estimation** *or* **determination of trend** *are as follows :*

1. Free hand, graphic method
2. Semi-average method
3. Moving average method
4. Method of least squares

GRAPHIC METHOD:

This is the simplest method of trend determination. According to this method, we plot the graph of the series and then draw a free hand curve through the points on the graph. Smoothing of time series data with a free hand curve eliminates the other components, viz., seasonal and irregular. The method does not involve complex mathematical calculations and can be used to describe all types of trend, linear or non-linear. However, the method is very subjective and can be adopted only to have a general idea of the nature of trend.

Example 1 *: Using the free hand hand or graphic method, fit a straight line trend to the following time series*

Year	1983	1984	1985	1986	1987	1988	1989	1990
Sales ('000)	80	90	85	92	87	99	93	120

Solution : Choosing a suitable scale, years are marked along the x-axis and corresponding sales values are marked along the y-axis. The points so obtained are then joined by straight lines which show the behaviour of sale values (actual data) over the given period. Then we draw a free hand straight line through the points of actual data for smoothing the time series data to obtain the trend. The behaviour of actual data and the trend line (dotted) are shown in fig. 1.

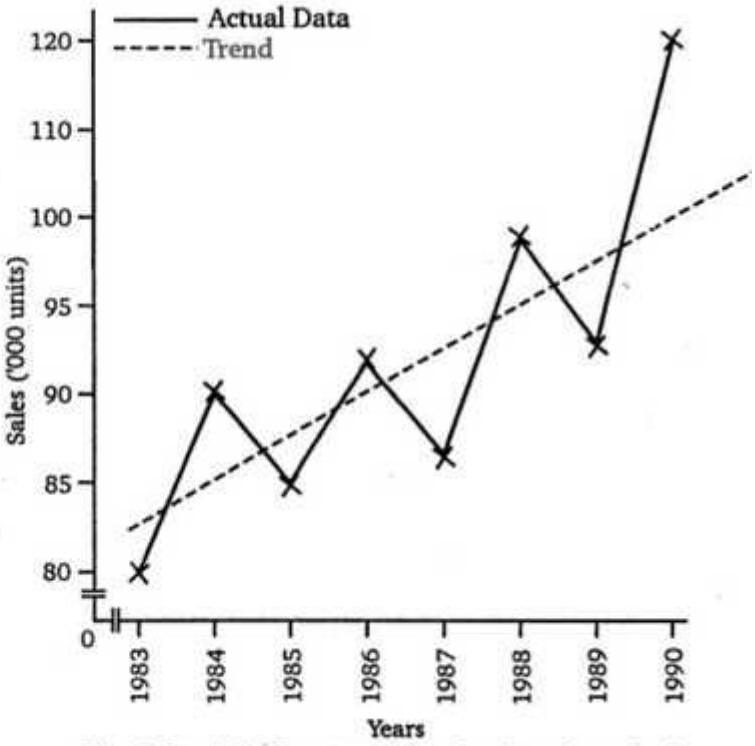


Fig. 1 Straight line trend by free hand method

SEMI- AVERAGE METHOD:

The method of semi-average is also simple. The method consists of dividing the data into two parts, preferably equal, and averaging the data in each part. In this way we obtain two points on the graph of the time series. The line obtained by joining these two points is the required trend line and may be extended in both the directions for estimating the trend values.

As compared with graphic method, the present method is better in view of its objectivity in the sense that every one who applies it would get the same results. However, the method has its limitation as it is appblicable only in a situation when the trend is linear or nearly linear. The following example will clarify the procedure.

Example 2 : *Determine straight line trend by semi-average method for the following time series data*

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production ('000 units)	18	25	21	15	26	31	30	20	35	32	23

Solution : According to semi-average method, the given time series is divided into two parts. Here,the data about 11 years are given, thus the value corresponding to the middle year, *i. e.*, 1985 is ignored. The averages of first and the last five years are then computed as under :

	Year	Production ('000 units)	Total Production	Semi average	Average year
First five years	1980	18			
	1981	25			
	1982	21	→ 105	105 ÷ 5 = 21	1982
	1983	15			
	1984	26			
Last five years	1986	30			
	1987	20			
	1988	35	→ 140	140 ÷ 5 = 28	1988
	1989	32			
	1990	23			

MOVING AVERAGE METHOD:

The method of moving averages attempts to smooth out the irregularities in a series by a process of averaging. *By using averages of appropriate orders (or extent), cyclical, seasonal and irregular variations may be eliminated, thus leaving only the trend component.* **Moving averages of extent** *m (or period) is a series of successive averages of m terms at a time, starting from 1st, 2nd, 3rd terms and so on until we exhaust the whole time series.* If m is odd, say equal to $(2k + 1)$, then the moving average is put against the mid-value of the period it covers, *i.e.*, against $t = k + 1$. On the other hand, if m is even, say equal to $2k$, it is placed between two middle values of the period it covers. Thus when an even number of years is taken in moving average, the average does not coincide with an original time period. For overcoming this situation, moving average of extent two of these moving averages are taken and the first of such values is put against $t = k + 1$. This procedure of centering puts the moving averages against the time points of the series rather than between these points. Symbolically, the 3-yearly moving averages of a time series can be computed as shown in the following table.

3 YEARLY MOVING AVERAGE			
Col. 1.	**Col. 2.**	**Col. 3.**	**Col. 4 = Col. 3 ÷ 3**
Years (t)	y_t	**3-yearly moving totals**	**3-yearly moving averages**
1	y_1	–	–
2	y_2	$\rightarrow (y_1 + y_2 + y_3)$	$(y_1 + y_2 + y_3)/3$
3	y_3	$\rightarrow (y_2 + y_3 + y_4)$	$(y_2 + y_3 + y_4)/3$
4	y_4	$\rightarrow (y_3 + y_4 + y_5)$	$(y_3 + y_4 + y_5)/3$
5	y_5	$\rightarrow (y_4 + y_5 + y_6)$	$(y_4 + y_5 + y_6)/3$
6	y_6	$\rightarrow (y_5 + y_6 + y_7)$	$(y_5 + y_6 + y_7)/3$
7	y_7	$\rightarrow (y_6 + y_7 + y_8)$	$(y_6 + y_7 + y_8)/3$
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
$N-1$	y_{N-1}	$\rightarrow (y_{N-2} + y_{N-1} + y_N)$	$(y_{N-2} + y_{N-1} + y_N)/3$
N	y_N	–	–

EXAMPLE 1:

Using three year moving averages determine the trend and short term fluctuations.

Year :	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Production: ('000 tons)	21	22	23	25	24	22	25	26	27	26

Solution:

year	production	3 year moving total	3 year moving average	Short term fluctuation
1973	21	...	...	...
1974	22	66	22.00	0.00
1975	23	70	23.33	-0.33
1976	25	72	24.00	1.00
1977	24	71	23.67	0.33
1978	22	71	23.67	-1.67
1979	25	73	24.33	0.67
1980	26	78	26.00	0.00
1981	27	79	26.33	0.67
1982	26	...	...	...

Example :2

Obtain trend for four yearly moving averages for the following data.

Year:	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Production:	614	615	652	678	681	655	717	719	708	779	757

Solution : **Computation of trend by 4-yearly moving averages**

Year	Production	4-yearly moving totals	4-yearly centred moving totals	4-yearly moving averages (trend)
(1)	(2)	(3)	(4)	Col. (4) ÷ 8
1988	614			–
1989	615			–
		→ 2559		
1990	652		→ 5185	648.125
		→ 2626		
1991	678		→ 5292	661.500
		→ 2666		
1992	681		→ 5397	674.625
		→ 2731		
1993	655		→ 5503	687.875
		→ 2772		
1994	717		→ 5571	696.375

		⟶	2799			
1995	719			⟶	5722	715.250
		⟶	2923			
1996	708			⟶	5886	735.750
		⟶	2963			
1997	779					–
1998	757					–

In this case the following steps are followed :

1. Calculate 4-yearly moving totals as usual. These are given in column (3).
2. For centring, we obtain two-yearly moving total of the 4-yearly moving totals as shown in column (4). Let us call such centred total as 4-yearly centred moving totals.
3. Finally divide the 4-yearly centred moving totals by 8 (4 × 2, *i.e.*, the period or extent of moving average × 2) to get the 4-yearly centred moving averages or Trend values.

METHOD OF LEAST SQUARES:

The method of least squares has already been explained in the context of regression analysis in chapter 10 of the present book. *As observed, the method is very useful for fitting mathematical functions to a given set of data.* The method is objective, and therefore, gives correct and accurate estimation of trend, once the form of equation representing trend is determined.

An examination of graphical plot of the time series often provides an adequate basis for deciding the functional form of the trend. Some of the common curves used for representing trend are :

(a) $Y = a + bX$, Linear or Straight line trend.

(b) $Y = a + bX + cX^2$, Parabolic or Quadratic trend.

(c) $Y = ab^X$, Exponential trend.

(a) Fitting of Linear or Straight Line Trend

The simplest type of trend equation is the linear equation of the form

$$Y = a + bX \qquad ...(1)$$

where X represents time and Y the value of the variable. Here Y is the dependent and X is an independent variable.

Now for the set of given data $(X_1, Y_1), (X_2, Y_2).....,(X_N, Y_N)$, the constants a and b are determined by solving simultaneously the equations :

$$\Sigma Y = Na + b \Sigma X$$

$$\Sigma XY = a \Sigma X + b \Sigma X^2 \qquad ...(2)$$

The equations in (2), called **normal equations for the least square line in (1),** gives

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{N\Sigma X^2 - (\Sigma X)^2} \quad ..(3)$$

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \quad ...(4)$$

If the values of X are equidistant, the calculations involved in the estimation of a and b can be further simplified by shifting the origin to the appropriate mid-point in time, so that $\Sigma X = 0$. Obviously, the normal equations in (2) becomes

$$\left.\begin{aligned} \Sigma Y &= Na \\ \Sigma XY &= b\Sigma X^2 \end{aligned}\right\} \quad ...(5)$$

Therefore, $a = \frac{\Sigma Y}{N}$ and $b = \frac{\Sigma XY}{\Sigma X^2}$ ...(6)

Substituting the estimated values of a and b in (1), the fitted linear trend will be

$$Y = a + bX \quad ...(7)$$

we can find the trend values, say Y, by putting different values of X in (7). When writing the trend equation, the origin and unit of time must be clearly specified, as an equation without such specification will be useless.

EXAMPLE:

Below are given the figures of production (in 1000 tons) of a fertilizer factory.

Year	1997	1998	1999	2000	2001	2002	2003
Production	70	75	90	98	84	91	99

Fit a straight line trend by the method os least squares and estimate trend values for 2005.

[U.P.T.U. 2008]

Solution **:** We use the method of least squares to fit a straight line trend. Here, the trend line is

$$Y = a + bX$$

where Y is the production

we make the transformation

$$x = X - 2000 \quad ...(i)$$

Thus, the trend becomes

$$Y = a + bx \quad ...(ii)$$

Computation of trend by least squares method

Year (X)	Number (Y)	$x = X - 2000$	x^2	xY
1997	70	−3	9	−210
1998	75	−2	4	−150
1999	90	−1	1	90
2000	98	0	0	0
2001	84	1	1	84
2002	91	2	4	182
2003	99	3	9	297
$N = 7$	$\Sigma Y = 607$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma xY = 113$

The normal equations are

$$\Sigma Y = N a + \Sigma X$$

$$\Sigma xY = a\Sigma X + b\Sigma x^2$$

From the table, these equations becomes

$$607 = 7a + 0 \quad \Rightarrow \quad a = 86.7$$

$$113 = 0 + 28b \quad \Rightarrow \quad b = 4.03$$

Thus, the fitted trend line becomes

$$Y = 86.7 + 4.03x \quad \text{where} \quad x = X - 2000 \qquad \ldots(iii)$$

Putting $x = -3, -2, -1, 0, 1, 2, 3$ in (iii) we can get trend values as follows :

Year	1997	1998	1999	2000	2001	2002	2003
Trend Values $Y = 86.7 + 4.03x$	74.61	78.64	82.67	86.7	90.73	94.76	98.79

Estimate of production for 2005 is

$$\hat{Y} = 86.7 + 4.03(2005 - 2000)$$
$$= 86.7 + 20.15$$
$$= 106.85$$

SEASONAL VARIATION:

As discussed earlier, *there are certain variations, called seasonal variations, which occur with certain degree of regularity within a definite period.* The period of variations may be a year, a month or even a day. A variety of causes may be listed for such variations. Some times climatic conditions affect production in agriculture and industries. For example, the sale of woollens picks up in every winter; prices of food grains come down in harvesting season; sale of cold drinks goes up during summer, etc. and so on. On the other hand, there are man-made factors which also cause such variations. For instance, the demand for consumer products goes up during the early part of month. The traffic in a city is high during the rush hours. When time series data are given in annual figures, it will not possess the seasonal variations. Thus, such variations are present only when data are given for specific periods of the year *i.e.*, the data are given quarterly, monthly, weekly, daily or hourly.

MEASURES OF SEASONAL VARIATION:

1. Method of averages
2. Moving Average Method
3. Ratio to moving average
4. Ratio to trend.

1. Method of Simple Averages

According to this method the data for each month (if monthly is given) are expressed as percentage of the average for the year. The method involves the following **steps** :

(i) Arrange the data by years and month (or quarters if quarterly data are given).

(ii) The figures for each month are added and averages are obtained by dividing the monthly totals by the number of years. Suppose the averages for the 12 months are denoted by $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{12}$.

(iii) Then obtain the overall average of monthly averages as :

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \ldots + \bar{X}_{12}}{12}$$

(v) Obtain **seasonal indices** for different months by expressing the monthly averages as percentages of the overall average $\bar{X}$ in the following way :

$$\text{Seasonal Index for the first month} = \frac{\bar{X}_1}{\bar{X}} \times 100$$

$$\text{Seasonal Index for the second month} = \frac{\bar{X}_2}{\bar{X}} \times 100$$

... ...

$$\text{Seasonal Index for the twelfth month} = \frac{\bar{X}_{12}}{\bar{X}} \times 100$$

It should be noted that the average of the indices will always be 100, *i. e.*, the sum of the indices will be 1200 for 12 monthly data and the sum will be 400 for 4 quarterly data.

Example:

Assuming that the trend is absent, determine if there is any seasonality in the data given below

Year	Ist Quarter	2nd Quarter	3rd Quarter	4th Quarter
2004	3.7	4.1	3.3	3.5
2005	3.7	3.9	3.6	3.6
2006	4.0	4.1	3.3	3.1
2007	3.3	4.4	4.0	4.0

What are the seasonal indices for various quarters ? *(M. Com., M.K. Univ.)*

Solution. COMPUTATION OF SEASONAL INDICES

Year	Ist Quarter	2nd Quarter	3rd Quarter	4th Quarter
2004	3.7	4.1	3.3	3.5
2005	3.7	3.9	3.6	3.6
2006	4.0	4.1	3.3	3.1
2007	3.3	4.4	4.0	4.0
Total	14.7	16.5	14.2	14.2
Average	3.675	4.125	3.55	3.55
Seasonal Index	98.66	110.74	95.30	95.30

Notes for calculating seasonal index

$$\text{The average of averages} = \frac{3.675 + 4.125 + 3.55 + 3.55}{4} = \frac{14.9}{4} = 3.725$$

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

$$\text{Seasonal Index for the first quarter} = \frac{3.675}{3.725} \times 100 = 98.66$$

$$\text{Seasonal Index for the second quarter} = \frac{4.125}{3.725} \times 100 = 110.74$$

$$\text{Seasonal Index for the third and fourth quarters} = \frac{3.55}{3.725} \times 100 = 95.30$$

2. Moving Average Method:

It is a method for computing trend values in a time series which eliminates the short term and random fluctuations from the time series by means of moving average. Moving average of a period m is a series of successive arithmetic means of m terms at a time starting with 1^{st}, 2^{nd}, 3^{rd} so on. The first average is the mean of first m terms; the second average is the mean of 2^{nd} term to (m+1)th term and 3^{rd} average is the mean of 3^{rd} term to (m+2)th term and so on. If m is odd then the moving average is placed against the mid value of the time interval it covers. But if m is even then the moving average lies between the two middle periods which does not correspond to any time period. So further steps has to be taken to place the moving average to a particular period of time. For that we take 2-yearly moving average of the moving averages which correspond to a particular time period. The resultant moving averages are the trend values.

Ex:1) Calculate 3-yearly moving average for the following data.

Years	Production	3-yearly moving avg (trend values)
1971-72	40	
1972-73	45	(40+45+40)/3 = 41.67
1973-74	40	(45+40+42)/3 = 42.33
1974-75	42	(40+42+46)/3 = 42.67
1975-76	46	(42+46+52)/3 = 46.67
1976-77	52	(46+52+56)/3 = 51.33
1977-78	56	(52+56+61)/3 = 56.33
1978-79	61	

Ex:1) Calculate 4-yearly moving average for the following data.

Years	Production	4-yearly moving avg	2-yealry moving avg (trend values)
1971-72	40		
1972-73	45		
		(40+45+40+42)/3 = 41.75	
1973-74	40		42.5
		(45+40+42+46)/3 = 43.15	
1974-75	42		44.12
		(40+42+46+52)/3 = 45	
1975-76	46		47
		(42+46+52+56)/3 = 49	
1976-77	52		51.38
		(46+52+56+61)/3 = 53.75	
1977-78	56		
1978-79	61		

3. Ratio to Trend Method:

Ratio-to-trend method is also known as **percentage trend method.** The method overcomes the difficulty of the simple average method when trend is present in the time series data. The method involves the following **steps** in measuring the seasonal indices :

(i) Compute the trend values by fitting trend equation to observed data by the method of least squares.

(ii) Express the original time series values as percentages of corresponding trend values.

(iii) Arrange these percentages according to years and months for monthly data (or according to years and quarters for quarterly data).

EXAMPLE:

The main defect of the ratio to trend method is that if there are cyclical swings in the series, the trend whether a straight line or a curve can never follow the actual data as closely as a 12- monthly moving average does. So a seasonal index computed by the ratio to moving average method may be less biased than the one calculated by the ratio to trend method.

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	30	40	36	34
2004	34	52	50	44
2005	40	58	54	48
2006	54	76	68	62
2007	80	92	86	82

Solution. For determining seasonal variation by ratio-to-trend method, first we will determine the trend for yearly data and then convert it to quarterly data.

CALCULATING TREND BY METHOD OF LEAST SQUARES

Year	*Yearly totals*	*Yearly average Y*	*Deviations from mid-year X*	*XY*	X^2	*Trend values*
2003	140	35	− 2	− 70	4	32
2004	180	45	− 1	− 45	1	44
2005	200	50	0	0	0	56
2006	260	65	+ 1	+ 65	1	68
2007	340	85	+ 2	+ 170	4	80
$N = 5$		$\Sigma Y = 280$		$\Sigma XY = 120$	$\Sigma X^2 = 10$	

The equation of the straight line trend is $Y = a + bX$.

$$a = \frac{\Sigma Y}{N} = \frac{280}{5} = 56 \qquad b = \frac{\Sigma XY}{\Sigma X^2} = \frac{120}{10} = 12$$

Quarterly increment $= \frac{12}{4} = 3$.

Calculation of Quarterly Trend Values. Consider 2003, trend value for the middle quarter, *i.e.*, half of 2nd and half of 3rd is 32. Quarterly increment is 3. So the trend value of 2nd quarter is $32 - \frac{3}{2}$, *i.e.*, 30.5 and for 3rd quarter is $32 + \frac{3}{2}$, *i.e.*, 33.5. Trend value for the 1st quarter is 30.5 – 3, *i.e.*, 27.5 and of 4th quarter is 33.5 + 3, *i.e.*, 36.5. We thus get quarterly trend values as shown below :

TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	27.5	30.5	33.5	36.5
2004	39.5	42.5	45.5	48.5
2005	51.5	54.5	57.5	60.5
2006	63.5	66.5	69.5	72.5
2007	75.5	78.5	81.5	84.5

The given values are expressed as percentage of the corresponding trend values.

Thus for 1st Qtr. of 2003, the percentage shall be $(30/27.5) \times 100 = 109.09$, for 2nd Qtr. $(40/30.5) \times 100 = 131.15$, etc.

GIVEN QUARTERLY VALUES AS % OF TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	109.09	131.15	107.46	93.15
2004	86.08	122.35	109.89	90.72
2005	77.67	106.42	93.91	79.34
2006	85.04	114.29	97.84	85.52
2007	105.96	117.20	105.52	97.04
Total	463.84	591.41	514.62	445.77
Average	92.77	118.28	102.92	89.15
S.I. Adjusted	92.05	117.36	102.12	88.46

Total of averages = 92.77 + 118.28 + 102.92 + 89.15 = 403.12.

Since the total is more than 400 an adjustment is made by multiplying each average by $\frac{400}{403.12}$ and final indices are obtained.

4. Ratio to moving average:

Ratio-to-moving average or percentage moving average method consists of expressing the original time series data as percentages of moving averages instead of percentages of trend values as in **'ratio-to-trend method'**, while rest of the steps are essentially the same. The procedure in this method consists of the following steps :

(i) Find the centred 12-monthly-moving averages (if monthly data are given) from the given time series data.

(ii) Express the original time series values as the percentage of the corresponding centred moving average values.

(iii) Average these percentages according to years and months and find averages over the years for all the 12 months.

(iv) Find the overall average of these 12-monthly averages. If the overall average is 100, the 12 monthly averages will be taken as seasonal indices, otherwise the monthly averages expressed as percentages of the overall average will be the required seasonal indices for the 12 months.

Symbolically, the logic behind the process may be explained as under :

The 12-monthly moving averages will eliminate the seasonal and irregular components and give us an estimate of the remaining two components namely trend (T) and cyclic (C). In multiplicative model we thus get an estimate of $T \times C$. Then the second step results in :

$$\frac{Y}{T \times C} \times 100 = \frac{T \times C \times S \times I}{T \times C} \times 100 = (S \times I) \times 100$$

Now on averaging over $S \times I$ in the third step, we are able to eliminate the irregular components with a possible bias. The final step givens us the adjusted seasonal indices.

Example 1:

Obtain seasonal indices by ratio to moving average method:

Year	Quarters			
	I	II	III	IV
2007	68	62	61	63
2008	65	58	66	61
2009	68	63	63	67

Solution : In the 'ratio-to-moving average' method, we first calculate 4 quarterly moving averages and ratios to moving averages as under :

Computation of Ratios to Moving Averages

Year and Quarter		Original data Y	4-quarterly moving totals	4-quarterly centred moving totals 4	4-quarterly centred moving averages (T)	Ratio to moving averages (percentage) = Y/T ×100
2007	I	68				
	II	62				
		→	254			
	III	61	→	505	63.125	96.63
		→	251			
	IV	63	→	498	62.250	101.20
		→	247			
2008	I	65	→	499	62.375	104.21
		→	252			
	II	58	→	502	62.750	92.43
		→	250			
	III	66	→	503	62.875	104.97
		→	253			
	IV	61	→	511	63.875	95.50
		→	258			
2009	I	68	→	513	64.125	106.04
		→	255			
	II	63	→	516	64.500	97.67
		→	261			
	III	63				
	IV	67				

Again, the percentage of original data to moving averages are arranged according to years and quarters to obtain the seasonal indices as shown in the following table :

Computation of Seasonal Indices

Year	Percentages to moving averages			
	I	II	III	IV
2007	–	–	96.63	101.20
2008	104.21	92.43	104.97	65.50
2009	106.04	97.67	–	–
Totals	210.25	190.10	201.60	196.70
Averages	105.125	95.05	100.80	98.35
Adjusted Quarterly Indices	$\frac{105.125}{99.83}\times 100$ = **105.30**	$\frac{95.05}{99.83}\times 100$ = **95.21**	$\frac{100.80}{99.83}\times 100$ = **100.97**	$\frac{98.35}{99.83}\times 100$ = **98.52**

Overall mean $= \bar{X} = \frac{105.125 + 95.05 + 100.80 + 98.35}{4} =$ **99.83**