

UNIT I BASIC STATISTICS – SMTA1207

Classification and Tabulation of Data

Classification of Data

The data that are unorganized or have not been arranged in any way are called raw data. The ungrouped data are often voluminous, complex to handle and hardly useful to draw any vital decisions. Hence, it is essential to rearrange the elements of the raw data set in a specific pattern. Further, it is important that such data must be presented in a condensed form and must be classified according to homogeneity for the purpose of analysis and interpretation. An arrangement of raw data in an order of magnitude or in a sequence is called **array**. Specifically, an arrangement of observations in an ascending or a descending order of magnitude is said to be an **ordered array**.

Classification is the process of arranging the primary data in a definite pattern and presenting in a systematic form. Horace Secrets defined classification as the process of arranging the data into sequences and groups according to their common characteristics or separating them into different but related parts. It is treated as the process of classifying the elements of observations or things into different groups or classes or sequences according to the resemblances and similarities of their character. It is also defined as the process of dividing the data into different groups or classes which are as homogeneous as possible within the groups or classes, but heterogeneous between themselves.

Objectives of Classification

Classification of data has manifold objectives. The salient features among them are the following:

- ☐ It explains the features of the data.
- ☐ It facilitates comparison with similar data.
- ☐ It strikes a note of homogeneity in the heterogeneous elements of the collected information.

- ☐ It explains the similarities which may exist in the diversity of data points.
- ☐ It is required to condense the mass data in such a manner that the similarities and dissimilarities are understood.
- ☐ It reduces the complexity of nature of data and renders the data to comprehend easily.
- ☐ It enables proper utilization of data for further statistical treatment.

Types of Classification

The raw data can be classified in various ways depending on the nature of data. The general types of classification are: (i) Classification by Time or Chronological Classification

(i) Classification by Space or Spatial Classification (iii) Classification by Attribute or Qualitative Classification and (iv) Classification by Size or Quantitative Classification. Each of these types is now described.

Classification by Time or Chronological Classification

The method of classifying data according to time component is known as classification by time or chronological classification. In this type of classification, the groups or classes are arranged either in the ascending order or in the descending order with reference to time such as years, quarters, months, weeks, days, etc. Illustrations for statistical data to be classified under this type are listed below:

- ☐ Number of new schools established in Tamil Nadu during 1995 – 2015
- ☐ Pass percentage of students in SSLC Board Examinations over a period of past
5 years
- ☐ Index of market prices in stock exchanges arranged day-wise
- ☐ Month-wise salary particulars of employees in an industry
- ☐ Particulars of outpatients in a Primary Health Centre presented day-wise.

The classification of data relating to the price of 10 gms of gold in India during 2001 - 2012

Table 3.1

Price of 10 gms of Gold in India

Year	Price in `	Year	Price in `	Year	Price in `
2001	4300	2005	7000	2009	14500
2002	4990	2006	8400	2010	18500
2003	5600	2007	10800	2011	26400
2004	5850	2008	12500	2012	31799

The classification of data relating to the population of India from 1961 to 2011 .

Table 3.2

Population of India from 1961 to 2011

Year	1961	1971	1981	1991	2001	2011
Population (in crores)	43.92	54.82	68.33	84.64	102.87	121.02

Classification by Space (Spatial) or Geographical Classification

The method of classifying data with reference to geographical location such as countries, states, cities, districts, etc., is called classification by space or spatial classification. It is also termed as geographical classification. The following are some examples:

- ☐ Number of school students in rural and urban areas in a State
- ☐ Region-wise literacy rate in a state
- ☐ State-wise crop production in India
- ☐ Country-wise growth rate in South East Asia

The classification of data relating to number of schools and types of schools in 7 major cities of Tamil Nadu as per the Annual Budget Report 2012 – 2013 is given in Table 3.3

Table 3.3
Number of Schools and Types of Schools

District	Primary School	Middle School	High School	Hr. Sec. School	Total
Chennai	697	203	206	448	1554
Coimbatore	1090	307	185	306	1888
Madurai	1314	332	172	254	2075
Trichy	1260	350	187	199	1996
Salem	1402	445	213	231	2291
Tirunelveli	1786	437	178	251	2652
Erode	986	357	146	176	1665

Average yield of rice (Kg/hect) during 2014-15 as per the records of Directorate of Economics and Statistics, Ministry of Agriculture and Farmers Welfare, Government of India, in five states in India is given in Table 3.4

Table 3.4
Average Yield of Rice during 2014 - 15

State	Yield (Kg/hect)
Tamilnadu	3191
Karnataka	2827
Kerala	2818
Uttarpradesh	2082
West Bengal	2731

Classification by Attributes or Qualitative classification

The method of classifying statistical data on the basis of attribute is said to be classification by attributes or qualitative classification. Examples of attributes include nationality, religion, gender, marital status, literacy and so on.

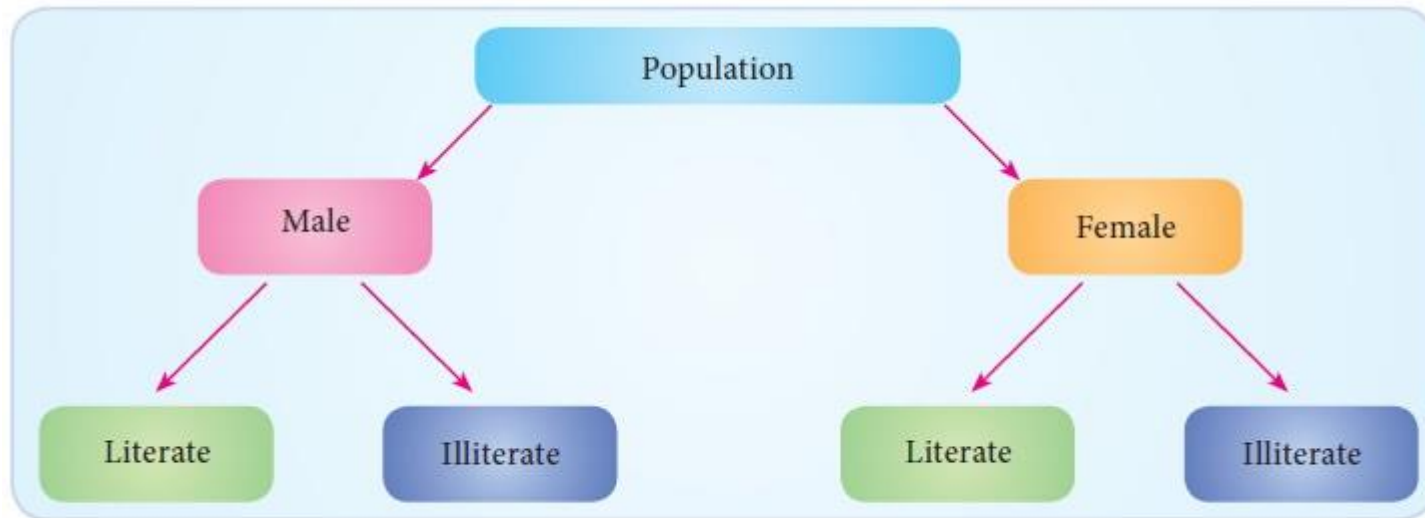


Fig: 3.1

Classification according to attributes is of two kinds: simple classification and manifold classification.

In simple classification the raw data are classified by a single attribute. All those units in which a particular characteristic is present are placed in one group and others are placed in another group. The classification of individuals according to literacy, gender, economic status would come under simple classification.

In manifold classification, two or more attributes are considered simultaneously. When more attributes are involved, the data would be classified into several classes and subclasses depending on the number of attributes. For example, population in a country can be classified in terms of gender as male and female. These two sub-classes may be further classified in terms of literacy as literate and illiterate.

While classifying the data according to attributes, it is essential to ensure that the attributes involved have to be defined without ambiguity. For example, while classifying income groups, the investigator has to define carefully the different non-overlapping income groups.

The classification of students studying in a school according to gender is given in Table 3. 5

Table 3.5

Gender-wise and class-wise information about students in a School

Class	Boys	Girls
VI	82	34
VII	74	43
VIII	92	27
IX	87	32
X	90	30
XI	75	25
XII	78	22

Classification by Size or Quantitative Classification

When the characteristics are measured on numerical scale, they may be classified on the basis of their magnitude. Such a classification is known as classification by size or quantitative classification. For example data relating to the characteristics such as height, weight, age, income, marks of students, production and consumption, etc., which are quantitative in nature,



Colours of vegetables,

Types of vegetables

Weight of vegetables,

Cost of vegetables

Qualitative data (Non-numerical)

Quantitative data (Numerical)

come under this category.

The classification of data relating to nutritive values of three items measured per 100 grams is provided in Table 3.6

Table 3.6
Nutritive values of Sugar, Jaggery and Honey

Item	Energy K calories	Carbohydrate (in gm)	Calcium (in mg)	Iron (in mg)
Sugar	398	99.4	12	0.15
Jaggery	383	95.0	80	2.65
Honey	313	79.5	5	0.69

Source: National Institute of Nutrition, ICMR, Hyderabad.

In the classification of data by size, data may also be classified deriving number of classes based on the range of observations and assigning number of observations lying in each class. The following is another example for classification by size.

The classification of 55 students according to their marks is given in Table 3.7

Table 3.7
Classification of students with respect to their marks

Marks	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30	30 - 35
Number of Students	2	6	13	17	11	4	2

Rules for Classification of Data

There are certain rules to be followed for classifying the data which are given below.

- ☐ The classes must be exhaustive, i.e., it should be possible to include each of the data points in one or the other group or class.
- ☐ The classes must be mutually exclusive, i.e., there should not be any overlapping.

- ☐ It must be ensured that number of classes should be neither too large or nor too small. Generally, the number of classes may be fixed between 4 and 15.
- ☐ The magnitude or width of all the classes should be equal in the entire classification.
- ☐ The system of open end classes may be avoided.

Tabulation

A logical step after classifying the statistical data is to present them in the form of tables. A table is a systematic organization of statistical data in rows and columns. The main objective of tabulation is to answer various queries concerning the investigation. Tables are very helpful while carrying out the analysis of collected data and subsequently for drawing inferences from them. It is considered as the final stage in the compilation of data and forms the basis for its further statistical treatment.

Advantages of Tabulation

- ☐ It is a logical step of presenting statistical data after classification.
- ☐ It enables the reader to understand the required information with ease as the information is contained in rows and columns with figures.
- ☐ It enables the investigator to present the data in a brief or condensed and compact form.
- ☐ Comparison is made simple by displaying data to be compared in a single table.
- ☐ It is easy to remember the data points or items if they are properly placed in the form of table, as it provides a kind of visual aid.
- ☐ It facilitates easy computation and helps easy detection of errors and omissions.
- ☐ It enables the reader to refer the data to be presented in a manner that suits for further statistical treatment and for making valid conclusions.

Types of Tables

Statistical tables can be classified under two general categories, namely, general tables and summary tables.

General tables contain a collection of detailed information including all that is relevant to the subject or theme. The main purpose of such tables is to present all the information available on a certain problem at one place for easy reference and they are usually placed in the appendices of reports.

Summary tables are designed to serve some specific purposes. They are smaller in size than general tables, emphasize on some aspect of data and are generally incorporated within the text. The summary tables are also called derivative tables because they are derived from the general tables. The information contained in the summary table aims at analysis and inference. Hence, they are also known as interpretative tables.

The statistical tables may further be classified into two broad classes namely simple tables and complex tables. A simple table summarizes information on a single characteristic and is also called a univariate tab

The marks secured by a batch of students in a class test are displayed in Table 3.8

Table 3.8
Marks of Students

Marks	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
Number of Students	10	12	17	20	15	6

This table is based on a single characteristic namely marks and from this table one may observe the number of students in each class of marks. The questions such as the number of students scored in the range 50 – 60, the maximum number of students in a specific range of marks and so on can be determined from this table.

A complex table summarizes the complicated information and presents them into two or more interrelated categories. For example, if there are two coordinate factors, the table is called a two-way table or bi-variate table; if the number of coordinate groups is three, it is a case of three-way tabulation, and if it is based on more than three coordinate groups, the table is known as higher order tabulation or a manifold tabulation.

Table 3.9 is an illustration for a two-way table, in which there are two characteristics, namely, marks secured by the students in the test and the gender of the students. The table provides information relating to two interrelated characteristics, such as marks and gender of students. It is observed from the table that 26 students have scored marks in the range 40 – 50 and among them students, 16 are males and 10 are females.

Table 3.9
Marks of Students

Marks	Number of Students		Total
	Males	Females	
30 – 40	8	6	14
40 – 50	16	10	26
50 – 60	14	16	30
60 - 70	12	8	20
70 – 80	6	4	10
Total	56	44	100

Table 3.10 is an example for a three – way table with three factors, namely, marks, gender and location.

Table 3.10
Marks of Students

Marks	Males		Total	Females		Total	Total		Total
	Urban	Rural		Urban	Rural		Urban	Rural	
30 – 40	4	4	8	4	2	6	8	6	14
40 – 50	10	6	16	5	5	10	15	11	26
50 – 60	8	6	14	9	7	16	17	13	30
60 - 70	7	5	12	5	3	8	12	8	20
70 – 80	5	1	6	2	2	4	7	3	10
Total	34	22	56	25	19	44	59	44	100

From this table, one may get information relating to the distribution of students according marks, gender and geographical location from where they hail.

Components of a Table

Generally a table should be comprised of the following components:

- i. Table number and title
- ii. Stub (the headings of rows)
- iii. Caption (the headings of columns)
- iv. Body of the table
- v. Foot notes
- vi. Sources of data.

i. Table Number and Title: Each table should be identified by a number given at the top. It should also have an appropriate short and self explanatory title indicating what exactly the table presents.

ii. Stub: Stubs stand for brief and self explanatory headings of rows.

iii. Caption: Caption stands for brief and self explanatory headings of columns. It may involve headings and sub-headings as well.

iv. Body of the Table: The body of the table should provide the numerical information in different cells.

v. Foot Note: The explanatory notes should be given as foot notes and must be complete in order to understand them at a later stage.

vi. Source of Data: It is always customary to provide source of data to enable the user to refer the original data. The source of data may be provided in a foot note at the bottom of the table.

A typical format of a table is given below:

Table Number		
Title of the Table		
Stub heading	Caption (Column headings)	Total
Stub (Row entries)	Body	
Total		
Foot note (if any)		
Source of Data (if any)		

General Precautions for Tabulation

The following points may be considered while constructing statistical tables:

- ☐ A table must be as precise as possible and easy to understand.
- ☐ It must be free from ambiguity so that main characteristics from the data can be easily brought out.
- ☐ Presenting a mass of data in a single table should be avoided. Displaying the data in a single table would increase the chances for occurrence of mistakes and would make the table unwieldy. Such data may be presented in more than one table such that each table should be complete and should serve the purpose.
- ☐ Figures presented in columns for comparison must be placed as near to each other as possible. Percentages, totals and averages must be kept close to each other. Totals to be compared may be given in bold type wherever necessary.
- ☐ Each table should have an appropriate short and self- explanatory title indicating what exactly the table presents.
- ☐ The main headings and subheadings must be properly placed.
- ☐ The source of the data must be indicated in the footnote.
- ☐ The explanatory notes should always be given as footnotes and must be complete in order to understand them at a later stage.
- ☐ The column or row heads should indicate the units of measurements such as monetary units like Rupees, and other units such as meters, etc. wherever necessary.
- ☐ Column heading may be numbered for comparison purposes. Items may be arranged either in the order of their magnitude or in alphabetical, geographical, and chronological or in any other suitable arrangement for meaningful presentation.
- ☐ Figures as accurate as possible are to be entered in a table. If the figures are approximate, the same may be properly indicated.

Frequency Distribution

A tabular arrangement of raw data by a certain number of classes and the number of items (called frequency) belonging to each class is termed as a frequency distribution. The frequency distributions are of two types, namely, discrete frequency distribution and continuous frequency distribution.

Discrete Frequency Distribution

Raw data sometimes may contain a limited number of values and each of them appeared many numbers of times. Such data may be organized in a tabular form termed as a simple frequency distribution. Thus the tabular arrangement of the data values along with the frequencies is a simple frequency distribution. A simple frequency distribution is formed using a tool called 'tally chart'. A tally chart is constructed using the following method:

- ☐ Examine each data value.
- ☐ Record the occurrence of the value with the slash symbol (/), called tally bar or tally mark.
- ☐ If the tally marks are more than four, put a crossbar on the four tally bar and make this as block of 5 tally bars (////)
- ☐ Find the frequency of the data value as the total number of tally bars i.e., tally marks corresponding to that value.

The marks obtained by 25 students in a test are given as follows: 10, 20, 20, 30, 40, 25, 25, 30, 40, 20, 25, 25, 50, 15, 25, 30, 40, 50, 40, 50, 30, 25, 25, 15 and 40. The following discrete frequency distribution represents the given data:

Table 3.11
Marks Scored by the Students

Marks	Tally Bars	Frequency
10	/	1
15	//	2
20	///	3
25	//// //	7
30	////	4
40	////	5
50	///	3
Total		25

Continuous Frequency Distribution:

It is necessary to summarize and present large masses of data so that important facts from the data could be extracted for effective decisions. A large mass of data that is summarized in such a way that the data values are distributed into groups, or classes, or categories along with the frequencies is known as a continuous or grouped frequency distribution.

Table 3.12 displays the number of orders for supply of machineries received by an industrial plant each week over a period of one year.

Table 3.12

Supply Orders for Machineries of an Industrial Plant

Number of Orders Received	Number of Weeks
0 – 4	2
5 – 9	8
10 – 14	11
15 – 19	14
20 – 24	6
25 – 29	4
30 – 34	3
35 – 39	2
40 – 44	1
45 – 49	1

This table is a grouped frequency distribution in which the number of orders are given as classes and number of weeks as frequencies. Some terminologies related to a frequency distribution are given below.

Class: If the observations of a data set are divided into groups and the groups are bounded by limits, then each group is called a class.

Class limits: The end values of a class are called class limits. The smaller value of the class limits is called lower limit (L) and the larger value is called the upper limit(U).

Class interval: The difference between the upper limit and the lower limit is called class interval (I). That is, $I = U - L$.

Class boundaries: Class boundaries are the midpoints between the upper limit of a class and the lower limit of its succeeding class in the sequence. Therefore, each class has an upper and lower boundaries.

Width: Width of a particular class is the difference between the upper class boundary and lower class boundary.

Mid-point: Half of the difference between the upper class boundary and lower class boundary.

In Example, the interval 0 - 4 is a class interval with 0 as the lower limit and 4 as the upper limit. The upper boundary of this class is obtained as midpoint of the upper limit of this class and lower limit of its succeeding class. Thus the upper boundary of the class 0 - 4 is 4.5. The lower class boundary of this is 0 - 0.5 which is - 0.5. The lower boundary of the class 5 - 9 is clearly 4.5. Similarly, the other boundaries of different classes can be found. The width of the classes is 5.

Inclusive and Exclusive Methods of Forming Frequency Distribution

Formation of frequency distribution is usually done by two different methods, namely inclusive method and exclusive method.

Inclusive method

In this method, both the lower and upper class limits are included in the classes. Inclusive type of classification may be used for a grouped frequency distribution for discrete variable like members in a family, number of workers etc., It cannot be used in the case of continuous variable like height, weight etc., where integral as well as fractional values are permissible. Since both upper limit and lower limit of classes are included for frequency calculation, this method is called inclusive method.

Exclusive method

In this method, the values which are equal to upper limit of a class are not included in that class and instead they would be included in the next class. The upper limit is not at all taken into consideration or in other words it is always excluded from the consideration. Hence this method is called exclusive method .

The marks scored by 50 students in an examination are given as follows:

23, 25, 36, 39, 37, 41, 42, 22, 26, 35, 34, 30, 29, 27, 47, 40, 31, 32, 43, 45, 34, 46, 23, 24, 27, 36, 41, 43, 39, 38, 28, 32, 42, 33, 46, 23, 34, 41, 40, 30, 45, 42, 39, 37, 38, 42, 44, 46, 29, 37.

It can be observed from this data set that the marks of 50 students vary from 22 to 47. If it is decided to divide this group into 6 smaller groups, we can have the boundary lines fixed as 25, 30, 35, 40, 45 and 50 marks. Then, we form the six groups with the boundaries as 21 - 25, 26 - 30, 31 - 35, 36 - 40, 41 - 45 and 46 - 50.

The continuous frequency distribution formed by inclusive and exclusive methods are displayed in Table 3.13(i) and Table 3.13(ii), respectively.

Table 3.13(i)

Marks secured by students (Inclusive Method)

Marks	x , Integer value	Tally Marks	No. of Students
21-25	$21 \leq x \leq 25$	 /	6
26-30	$26 \leq x \leq 30$	 	8
31-35	$31 \leq x \leq 35$	 	8
36-40	$36 \leq x \leq 40$	 //	12
41-45	$41 \leq x \leq 45$	 //	12
46-50	$46 \leq x \leq 50$		4
Total			50

Table 3.13(ii)

Marks secured by students (Exclusive Method)

Marks	x , Integer value	Tally Marks	No. of Students
20-25	$20 \leq x < 25$	 	5
25-30	$25 \leq x < 30$	 //	7
30-35	$30 \leq x < 35$	 	9
35-40	$35 \leq x < 40$	 /	11
40-45	$40 \leq x < 45$	 //	12
45-50	$45 \leq x < 50$	 /	6
Total			50

True class intervals

In the case of continuous variables, we take the classes in such a way that there is no gap between successive classes. The classes are defined in such a way that the upper limit of each class is equal to lower limit of the succeeding class. Such classes are known as true classes. The inclusive method of forming class intervals are also known as not-true classes. We can convert the not-true classes into true-classes by subtracting 0.5 from the lower limit of the class and adding 0.5 to the upper limit of each class like 19.5 - 25.5, 25.5 - 30.5, 30.5 - 35.5, 35.5 - 40.5, 40.5 - 45.5, 45.5 - 50.5.

Open End Classes

When a class limit is missing either at the lower end of the first class interval or at the upper end of the last classes or when the limits are not specified at both the ends, the frequency distribution is said to be the frequency distribution with open end classes.

Salary received by 113 workers in a factory are classified into 6 classes. The classes and their frequencies are displayed in Table 3.14 Since the lower limit of the first class and the upper limit of the last class are not specified, they are open end classes.

Table 3.14

Open-Ended Frequency Table

Salary Range in Rs.	Number of workers
Below 10000	18
10000 - 20000	23
20000 - 30000	30
30000 - 40000	20
40000 - 50000	12
50000 and above	10

Guidelines on Compilation of Continuous Frequency Distribution

The following guidelines may be followed for compiling the continuous frequency distribution.

- ☐ The values given in the data set must be contained within one (and only one) class and overlapping classes must not occur.

- The classes must be arranged in the order of their magnitude.
- Normally a frequency distribution may have 8 to 10 classes. It is not desirable to have less than 5 and more than 15 classes.
- Frequency distributions having equal class widths throughout are preferable. When this is not possible, classes with smaller or larger widths can be used. Open ended classes are acceptable but only in the first and the last classes of the distribution.
- It should be noted that in a frequency distribution, the first class should contain the lowest value and the last class should contain the highest value.
- The number of classes may be determined by using the Sturges formula $k = 1 + 3.322 \log_{10} N$, where N is the total frequency and k is the number of classes.

Cumulative Frequency Distribution

Cumulative frequency corresponding to a class interval is defined as the total frequency of all values less than upper boundary of the class. A tabular arrangement of all cumulative frequencies together with the corresponding classes is called a cumulative frequency distribution or cumulative frequency table.

The main difference between a frequency distribution and a cumulative frequency distribution is that in the former case a particular class interval according to how many items lie within it is described, whereas in the latter case the number of items that have values either above or below a particular level is described.

There are two forms of cumulative frequency distributions, which are defined as follows:

(i) **Less than Cumulative Frequency Distribution:** In this type of cumulative frequency distribution, the cumulative frequency for each class shows the number of elements in the data whose magnitudes are less than the upper limit of the respective class.

(ii) **More than Cumulative Frequency Distribution:** In this type of cumulative frequency distribution, the cumulative frequency for each class shows the number of elements in the data whose magnitudes are larger than the lower limit of the class.

Construct less than and more than cumulative frequency distribution tables for the following frequency distribution of orders received by a business firm over a number of weeks during a year.

Number of order received	0 - 4	5 - 9	10 - 14	15 - 19	20 - 24
Number of weeks	2	8	11	14	6
Number of order received	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49
Number of weeks	4	3	2	1	1

Solution:

For the data related to the number of orders received per week by a business firm during a period of one year, the less than and more than cumulative frequencies are computed and are given in Table 3.15

Table 3.15

Cumulative Frequency Distribution for the number of orders received by a Business firm

Given data		Less than ogive		More than ogive	
Number of Orders Received	Number of Weeks	Upper limit	Less than Cumulative Frequencies	Lower limit	More Than Cumulative Frequencies
0 - 4	2	4	2	0	52
5 - 9	8	9	10	5	50
10 - 14	11	14	21	10	42
15 - 19	14	19	35	15	31
20 - 24	6	24	41	20	17
25 - 29	4	29	45	25	11
30 - 34	3	34	48	30	7
35 - 39	2	39	50	35	4
40 - 44	1	44	51	40	2
45 - 49	1	49	52	45	1

Relative-Cumulative Frequency Distributions

The relative cumulative frequency is defined as the ratio of the cumulative frequency to the total frequency. The relative cumulative frequency is usually expressed in terms of a percentage. The arrangement of relative cumulative frequencies against the respective class boundaries is termed as relative cumulative frequency distribution or percentage cumulative frequency distribution.

Bivariate Frequency Distributions

It is known that the frequency distribution of a single variable is called univariate distribution. When a data set consists of a large mass of observations, they may be summarized by using a two-way table. A two-way table is associated with two variables, say X and Y. For each variable, a number of classes can be defined keeping in view the same considerations as in the univariate case. When there are m classes for X and n classes for Y, there will be $m \times n$ cells in the two-way table. The classes of one variable may be arranged horizontally, and the classes of another variable may be arranged vertically in the two way table. By going through the pairs of values of X and Y, we can find the frequency for each cell. The whole set of cell frequencies will then define a bivariate frequency distribution. In other words, a bivariate frequency distribution is the frequency distribution of two variables.

Table 3.17 shows the frequency distribution of two variables, namely, age and marks obtained by 50 students in an intelligent test. Classes defined for marks are arranged horizontally (rows) and the classes defined for age are arranged vertically (columns). Each cell shows the frequency of the corresponding row and column values. For instance, there are 5 students whose age fall in the class 20 – 22 years and their marks lie in the group 30 – 40.

Table 3.17

Bivariate Frequency Distribution of Age and Marks

Marks	Age in Years				Total
	16 – 18	18 - 20	20 - 22	22 – 24	
10 – 20	2	1	1	-	4
20 – 30	3	2	3	1	9
30 – 40	3	3	5	6	17
40 – 50	2	2	3	4	11
50 – 60	-	1	2	2	5
60 – 70	-	1	2	1	4
Total	10	10	16	14	50

Meaning and Significance of Diagrams and Graphs

Diagrams:

A diagram is a visual form for presenting statistical data for highlighting the basic facts and relationship which are inherent in the data. The diagrammatic presentation is more understandable and it is appreciated by everyone. It attracts the attention and it is a quicker way of grasping the results saving the time. It is very much required, particularly, in presenting qualitative data.

Graphs:

The quantitative data is usually represented by graphs. Though it is not quite attractive and understandable by a layman, the classification and tabulation techniques will reduce the complexity of presenting the data using graphs. Statisticians have understood the importance of graphical presentation to present the data in an interpretable way. The graphs are drawn manually on graph papers.

Significance of Diagrams and Graphs:

Diagrams and graphs are extremely useful due to the following reasons:

- ☐ They are attractive and impressive
- ☐ They make data more simple and intelligible
- ☐ They are amenable for comparison
- ☐ They save time and labour and
- ☐ They have great memorizing effect.

Rules for Constructing Diagrams

While constructing diagrams for statistical data, the following guidelines are to be kept in mind:

- ☐ A diagram should be neatly drawn in an attractive manner
- ☐ Every diagram must have a precise and suitable heading
- ☐ Appropriate scale has to be defined to present the diagram as per the size of the paper
- ☐ The scale should be mentioned in the diagram
- ☐ Mention the values of the independent variable along the X-axis and the values of the dependent variable along the Y-axis

- ☐ False base line(s) may be used in X-axis and Y-axis, if required
- ☐ Legends should be given for X-axis, Y-axis and each category of the independent variable to show the difference
- ☐ Foot notes can be given at the bottom of the diagram, if necessary

Types of Diagrams

In practice, varieties of diagrams are used to present the data. They are explained below.

Simple Bar Diagram

Simple bar diagram can be drawn either on horizontal or vertical base. But, bars on vertical base are more common. Bars are erected along the axis with uniform width and space between the bars must be equal. While constructing a simple bar diagram, the scale is determined as proportional to the highest value of the variable. The bars can be coloured to make the diagram attractive. This diagram is mostly drawn for categorical variable. It is more useful to present the data related to the fields of Business and Economics.

The production cost of the company in lakhs of rupees is given below.

- i. Construct a simple bar diagram.
- ii. Find in which year the production cost of the company is (a) maximum (b) minimum (c) less than 40 lakhs.
- iii. What is the average production cost of the company?
- iv. What is the percentage increase from 2014 to 2015?

Year	Production Cost
2010	55
2011	40
2012	30
2013	25
2014	35
2015	70

Solution:

(i) We represent the above data by simple bar diagram in the following manner:

Step-1: Years are marked along the X-axis and labelled as 'Year'.

Step-2: Values of Production Cost are marked along the Y-axis and labelled as 'Production Cost (in lakhs of `)'.
(in lakhs of `).

Step-3: Vertical rectangular bars are erected on the years marked and whose height is proportional to the magnitude of the respective production cost.

Step-4: Vertical bars are filled with the same colours.

The simple bar diagram is presented in Fig.4.1.



Fig 4.1

(ii) (a) The maximum production cost of the company was in the year 2015.

(b) The minimum production cost of the company was in the year 2013.

(c) The production cost of the company during the period 2012- 2014 is less than 40 lakhs.

iii. Average production Cost of the company

$$= \frac{55 + 40 + 30 + 25 + 35 + 70}{6}$$

$$= ₹ 42.5 \text{ Lakhs}$$

iv. Percentage increase in the production cost of the company is

$$= \frac{70}{35} \times 100$$

$$= 200\%$$

Pareto Diagram:

Vilfredo Pareto (1848-1923), born in Paris in an Italian aristocratic family, studied Engineering and Mathematics at the University of Turin. During his studies at the University of Lousane in Switzerland, Pareto derived a complicated mathematical formula to prove the distribution of income and wealth in society is not random. Approximately 80% of total wealth in a society lies with only 20% of the families. The famous law about the 'Vital few and trivial many' is widely known as 'Pareto Principle' in Economics.



Pareto diagram is similar to simple bar diagram. But, in Pareto diagram, the bars are arranged in the descending order of the heights of the bars. In addition, there will be a line representing the cumulative frequencies (in %) of the different categories of the variable. The line is more useful to find the vital categories among trivial category.

Administration of a school wished to initiate suitable preventive measures against breakage of equipment in its Chemistry laboratory. Information collected about breakage of equipment occurred during the year 2017 in the laboratory are given below:

Equipment	No. of breakages
Burette	45
Conical flask	75
Test tube	150
Pipette	30

Draw Pareto Diagram for the above data. Which equipment requires more attention in order to reduce breakages?

Solution:

Since we have to find the vital few among the several, we draw Pareto diagram.

Step 1 : Arrange the equipment according to the descending order of the number of breakages.

Step 2 : Find the percentage of breakages for each equipment using the formula *No. of Breakages*

$$= \frac{\text{No. of Breakages}}{\text{Total No. of Breakages}} \times 100$$

Step 3 : Calculate cumulative percentage for each equipment.

Step 4 : Mark the equipment along the X-axis and the number of breakages along the Y-axis. Construct an attached simple bar diagram to this data. In an attached simple bar diagram, the vertical bars are erected adjacently.

Step 5 : Mark the cumulative no. of breakages for each equipment corresponding to the mid-point of the respective vertical bar.

Step 6 : Draw a free hand curve joining those plotted points.

Equipment	No. of Breakages (Frequency)	No. of Breakages in percentage	Cumulative No. of Breakages in percentage
Test tube	150	$\frac{150}{300} \times 100 = 50$	50
Conical flask	75	$\frac{75}{300} \times 100 = 25$	75
Burette	45	$\frac{45}{300} \times 100 = 15$	90
Pipette	30	$\frac{30}{300} \times 100 = 10$	100
Total	300	100	

No of breakages in the chemistry laboratory

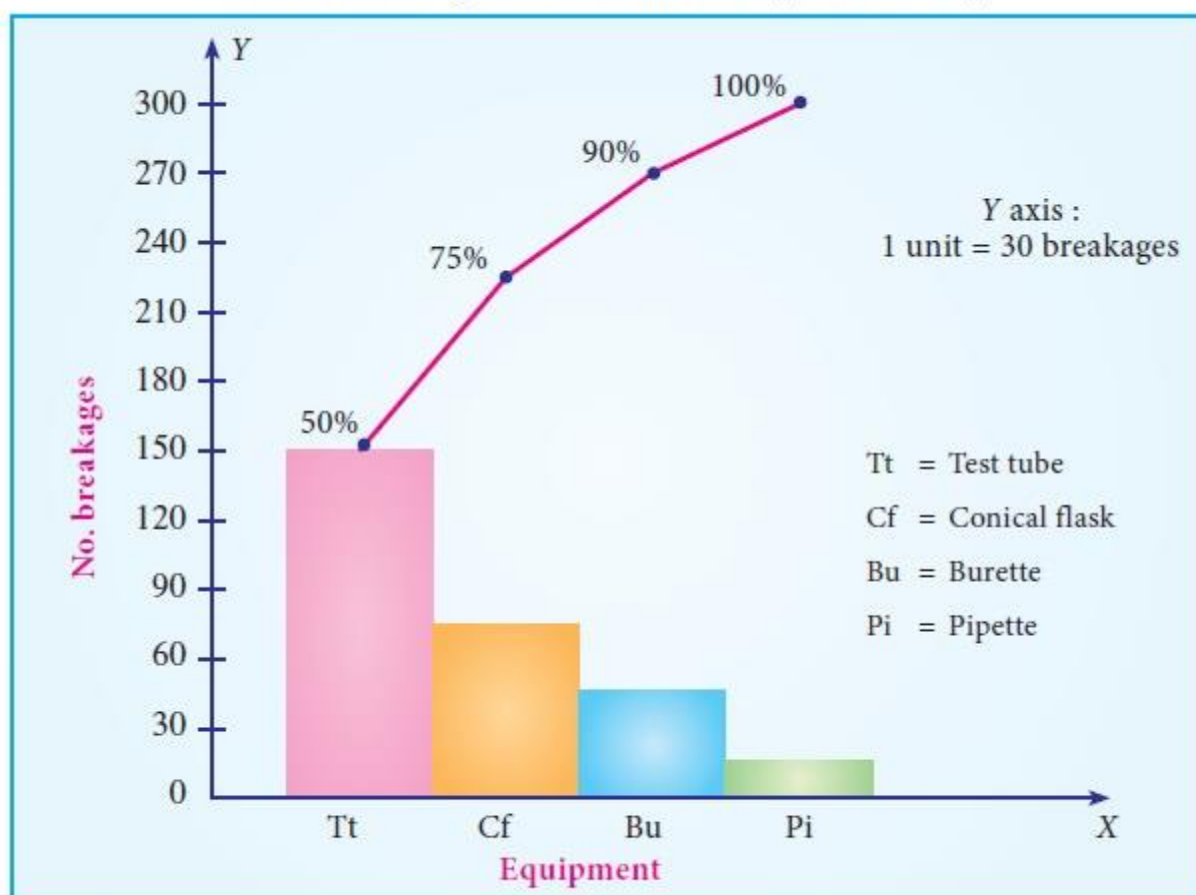


Fig 4.2: Pareto Diagram for No. of Breakages in the Chemistry Laboratory

From Fig 4.2, it can be found that 50% of breakages is due to Test tube, 25% due to Conical Flask. Therefore, the School Administration has to focus more attention on reducing the breakages of Test Tubes and Conical Flasks.

Multiple Bar Diagram

Multiple bar diagram is used for comparing two or more sets of statistical data. Bars with equal width are placed adjacently for each cluster of values of the variable. There should be equal space between clusters. In order to distinguish bars in each cluster, they may be either differently coloured or shaded. Legends should be provided

The table given below shows the profit obtained before and after tax payment(in lakhs of rupees) by a business man on selling cars from the year 2014 to 2017.

Year	Profit before tax	Profit after tax
2014	195	80
2015	200	87
2016	165	45
2017	140	32

- Construct a multiple bar diagram for the above data.
- In which year, the company earned maximum profit before paying the tax?
- In which year, the company earned minimum profit after paying the tax?
- Find the difference between the average profit earned by the company before paying the tax and after paying the tax.

Solution:

Since we are comparing the profit earned before and after paying the tax by the same Company, the multiple bar diagram is drawn. The diagram is drawn following the procedure presented below:

Step 1 : Years are marked along the X-axis and labeled as “Year”.

Step 2 : Values of Profit before and after paying the tax are marked along the Y-axis and labeled as “Profit (in lakhs of `)”.

Step 3 : Vertical rectangular bars are erected on the years marked, whose heights are proportional to the respective profit. The vertical bars corresponding to the profit earned before and after paying the tax in each year are placed adjacently.

Step 4 : The vertical bars drawn corresponding to the profit earned before paying the tax are filled with one type of colour. The vertical bars drawn corresponding to the profit earned after paying the tax are filled with another type of colour. The colouring procedure should be applied to all the years uniformly.

Step 5 : Legends are displayed to describe the different colours applied to the bars drawn for profit earned before and after paying the tax.

The multiple bar diagram is presented in Fig 4.3.

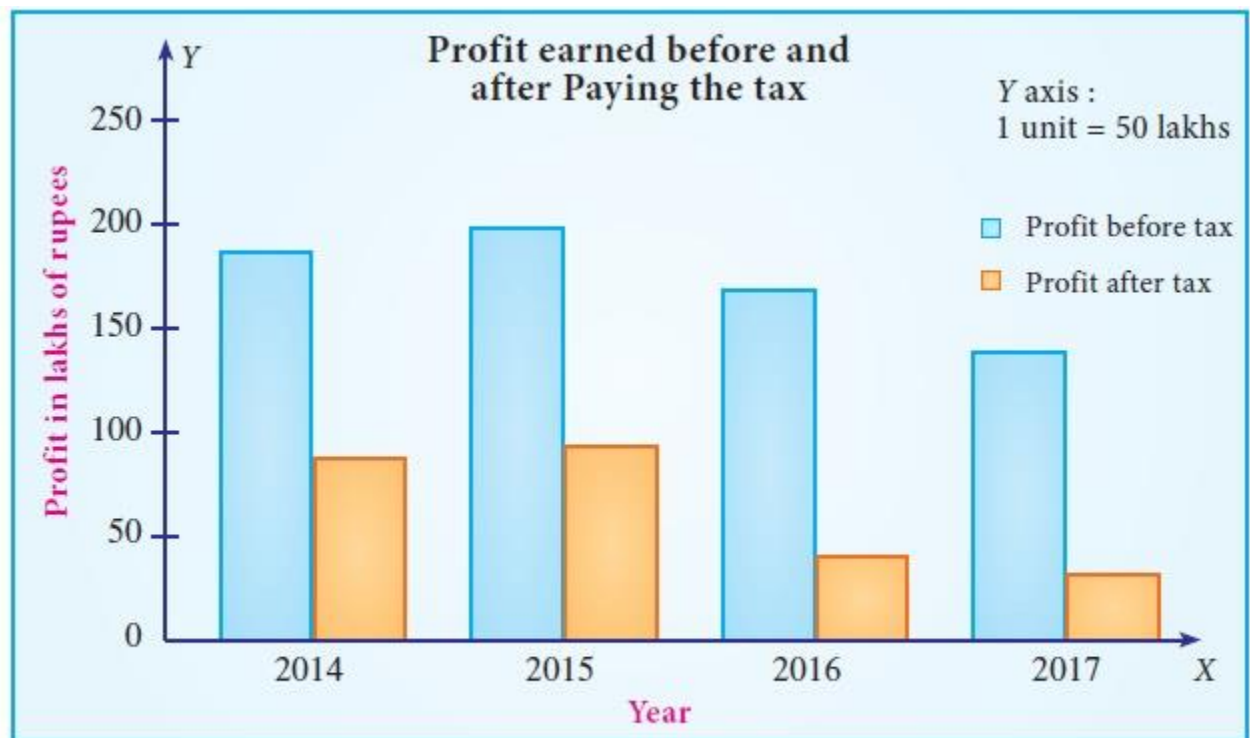


Fig 4.3 Multiple Bar Diagram for Profit by the Company earned before and after paying the Tax

- (i) The company earned the maximum profit before paying the tax in the year 2015.
- (ii) The company earned the minimum profit after paying the tax in the year 2017.
- (iii) The average profit earned before paying the tax = $700/4 = 175$ lak

The average profit earned after paying the tax = $244/4 = 61$ lakhs

Hence, difference between the average profit earned by the company before paying the tax and after paying the tax is = $175 - 61 = 114$ lakhs.

Component Bar Diagram(Sub-divided Bar Diagram)

A component bar diagram is used for comparing two or more sets of statistical data, as like multiple bar diagram. But, unlike multiple bar diagram, the bars are stacked in component bar diagrams. In the construction of sub-divided bar diagram, bars are drawn with equal width such that the heights of the bars are proportional to the magnitude of the total frequency. The bars are positioned with equal space. Each bar is sub-divided into various parts in proportion to the values of the components. The subdivisions are distinguished by different colours or shades. If the number of clusters and the categories in the clusters are large, the multiple bar diagram is not attractive due to more number of bars. In such situation, component bar diagram is preferred.

Total expenditure incurred on various heads of two schools in an year are given below. Draw a suitable bar diagram.

Expenditure Head	Amount (in lakhs)	
	School I	School II
Construction/Repairs	80	90
Computers	35	50
Laboratory	30	25
Watering plants	45	40
Library books	40	30
Total	230	235

Which school had spent more amount for

(a) construction/repairs (b) Watering plants?

Solution :

Since we are comparing the amount spent by two schools in a year towards various expenditures with respect to their total expenditures, a component bar diagram is drawn.

Step 1 : Schools are marked along the X-axis and labeled as “School”.

Step 2 : Expenditure Head are marked along the Y-axis and labeled as “Expenditure (in lakhs)”.

Step 3 : Vertical rectangular bars are erected for each school, whose heights are proportional to their respective total expenditure.

Step 4 : Each vertical bar is split into components in the order of the list of expenditure heads. Area of each rectangular box is proportional to the frequency of the respective expenditure

head/component. Rectangular boxes for each school are coloured with different colours. Same colours are applied to the similar expenditure heads for each school.

Step 5 : Legends are displayed to describe the colours applied to the rectangular boxes drawn for various expenditure heads.

The component bar diagram is presented in Fig 4.4.

Expenditure Head	Amount (₹ in lakhs)			
	School I		School II	
	Amount Spent	Cumulative Amount Spent	Amount Spent	Cumulative Amount Spent
Construction/Repairs	80	80	90	90
Computers	35	115	50	140
Laboratory	30	145	25	165
Watering plants	45	190	40	205
Library books	40	230	30	235

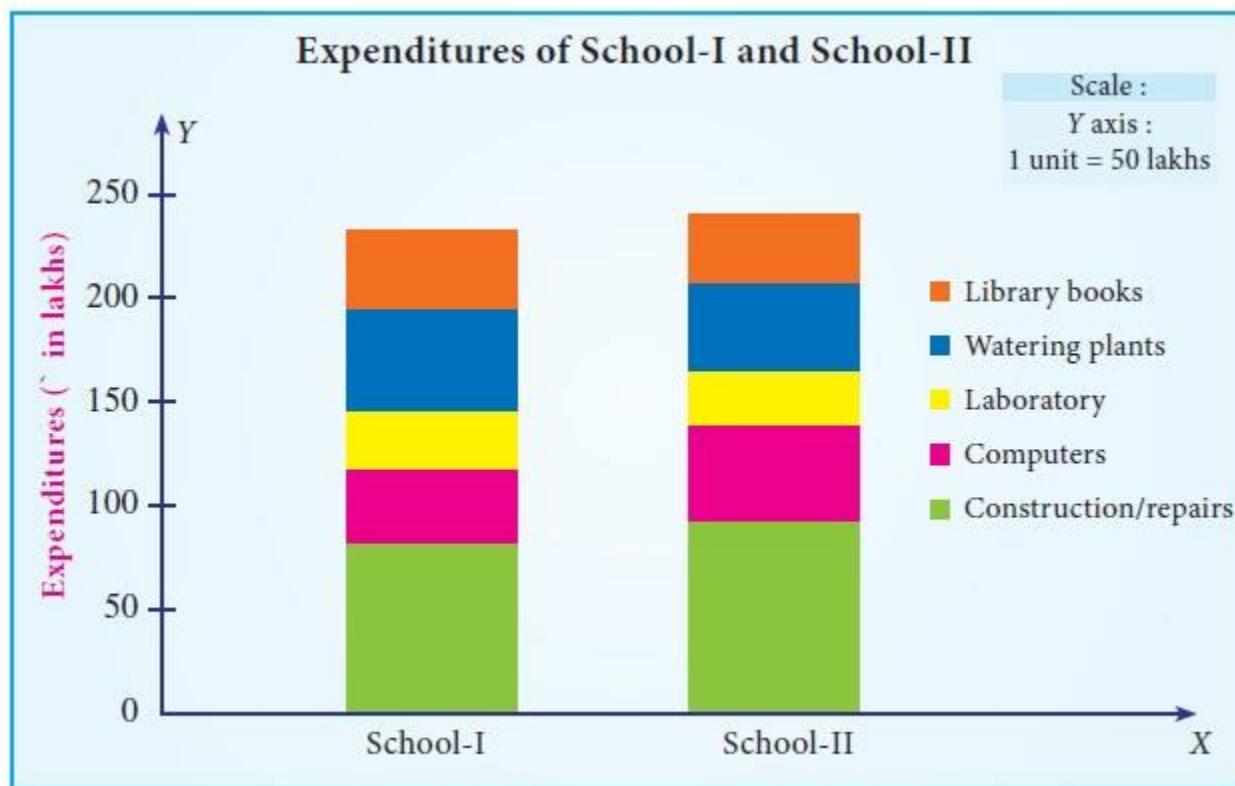


Fig 4.4 Component Bar diagram for expenditures of School I and School II

- (i) School- II had spent more amount towards Construction/Repairs.
- (ii) School- I had spent more amount towards Watering plants.

Percentage Bar Diagram

Percentage bar diagram is another form of component bar diagram. Here, the heights of the components do not represent the actual values, but percentages. The main difference between sub-divided bar diagram and percentage bar diagram is that, in the former, the height of the bars corresponds to the magnitude of the value. But, in the latter, it corresponds to the percentages. Thus, in the component bar diagram, heights of the bars are different, whereas in the percentage bar diagram, heights are equal corresponding to 100%. Hence, percentage bar diagram will be more appealing than sub-divided bar diagram. Also, comparison between components is much easier using percentage bar diagram.

Draw the percentage sub-divided bar diagram to the data given in Example 4.4. Also find

- (i) The percentage of amount spent for computers in School I
- (ii) What are the expenditures in which School II spent more than School.

Solution:

Since we are comparing the amount spent by two schools in a year towards various expenditures with respect to their total expenditures in percentages, a percentage bar diagram is drawn.

Step 1 : Schools are marked along the X-axis and labeled as “School”.

Step 2 : Amount spent in percentages are marked along the Y-axis and labeled as “Percentage of Expenditure (` in lakhs)”.

Step 3 : Vertical rectangular bars are erected for each school, whose heights are taken to be hundred.

Step 4 : Each vertical bar is split into components in the order of the list of percentage expenditure heads. Area of each rectangular box is proportional to the percentage of frequency of the respective expenditure head/component. Rectangular boxes for each school are coloured with different colours. Same colours are applied to the similar expenditure heads for each school.

Step 5 : Legends are displayed to describe the colours applied to the rectangular boxes drawn for various expenditure heads.

The percentage bar diagram is presented in Fig 4.5.

Expenditure Head	Amount (₹ in lakhs)					
	School I			School II		
	Amount spent	Percentage of Amount spent	Cumulative Percentage	Amount spent	Percentage of Amount spent	Cumulative Percentage
Construction / Repairs	80	35	35	90	38	38
Computers	35	15	50	50	21	59
Laboratory	30	13	63	25	11	70
Watering plants	45	20	83	40	17	87
Library books	40	17	100	30	13	100
Total	230	100		235	100	

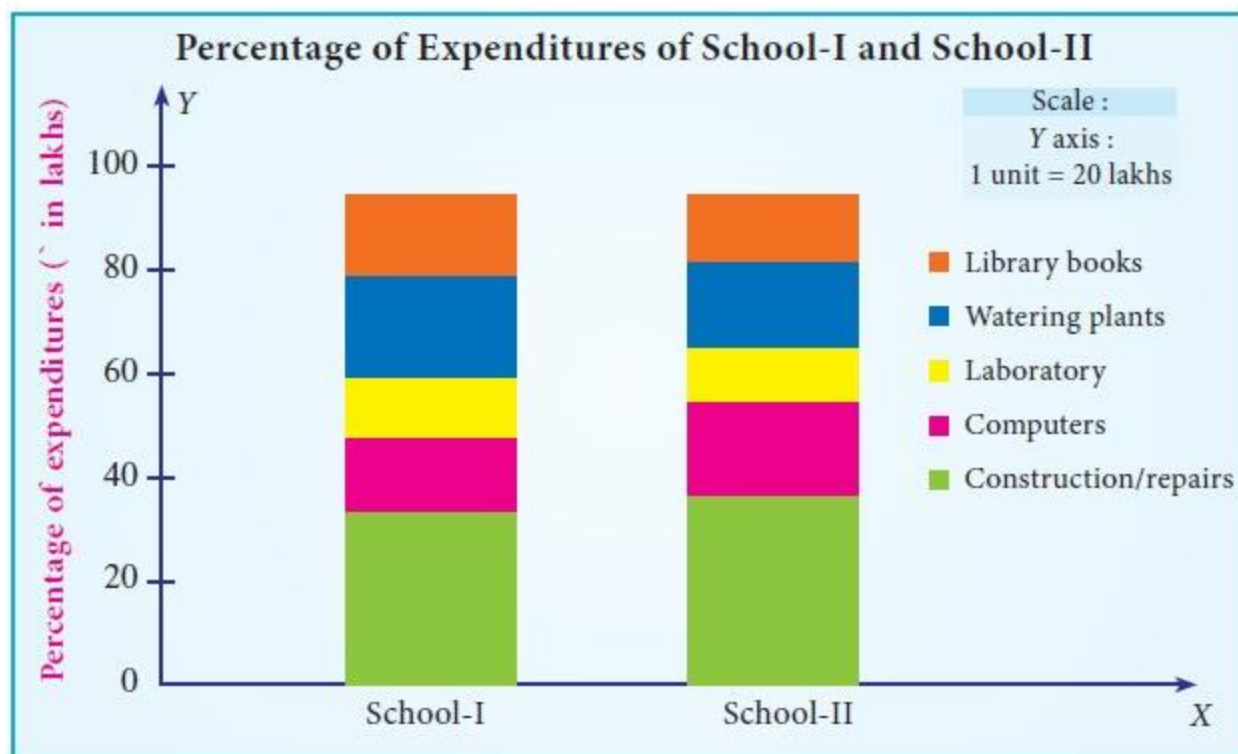


Fig 4.5 Percentage Bar diagram for expenditures of School I and School II

- 21% of the amount was spent for computers in School I
- 38% of expenditure was spent for construction/Repairs by School II than School I.

Pie Diagram

The Pie diagram is a circular diagram. As the diagram looks like a pie, it is given this name. A circle which has 360° is divided into different sectors. Angles of the sectors, subtending at the center, are proportional to the magnitudes of the frequency of the components.

Procedure:

The following procedure can be followed to draw a Pie diagram for a given data:

- Calculate total frequency, say, N.

- ii. Compute angles for each component using the formula.

$$\frac{\text{class frequency}}{N} \times 360$$

- iii. Draw a circle with radius of sufficient length as a horizontal line.
- iv. Draw the first sector in the anti-clockwise direction at an angle calculated for the first component.
- v. Draw the second sector adjacent to the first sector at an angle corresponding to the second component.
- vi. This process may be continued for all the components.
- vii. Shade/colour each sector with different shades/colours.
- viii. Write legends to each component.

Draw a pie diagram for the following data (in hundreds) of house hold expenditure of a family.

Items	Expenditure
Food	87
Clothing	24
Recreation	11
Education	13
Rent	25
Miscellaneous	20

Also find

- i. The central angle of the sector corresponding to the expenditure incurred on Education
- ii. By how much percentage the recreation cost is less than the Rent.

Solution :

The following procedure is followed to draw a Pie diagram for a given data:

- i. Calculate the total expenditure, say, N.
- ii. Compute angles for each component food, clothing, recreation, education, rent and miscellaneous using the formula $\frac{\text{class frequency}}{N} \times 360$

$$\frac{\text{class frequency}}{N} \times 360$$

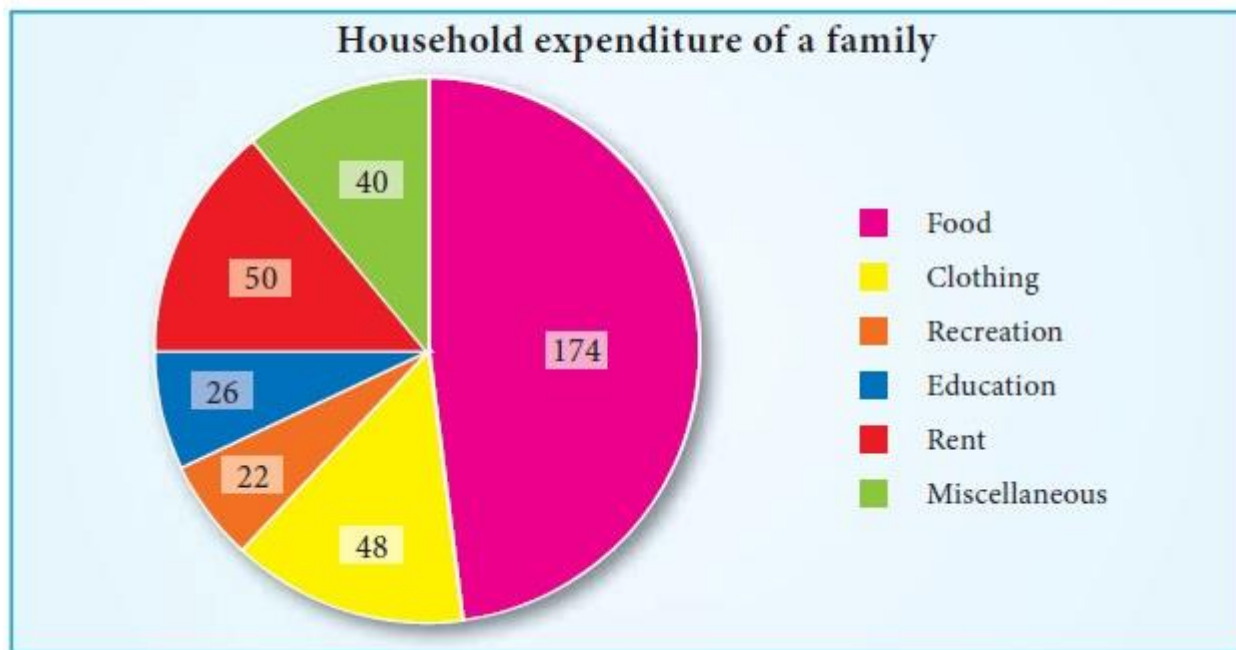
Item	Expenditure	Angle of the circle
Food	87	$\frac{87}{180} \times 360 = 174$
Clothing	24	$\frac{24}{180} \times 360 = 48$
Recreation	11	$\frac{11}{180} \times 360 = 22$
Education	13	$\frac{13}{180} \times 360 = 26$
Rent	25	$\frac{25}{180} \times 360 = 50$
Miscellaneous	20	$\frac{20}{180} \times 360 = 40$
Total	N=180	360

- iii. Draw a circle with radius of sufficient length as a horizontal line.
- iv. Draw the first sector in the anti-clockwise direction at an angle calculated for the first component food.
- v. Draw the second sector adjacent to the first sector at an angle corresponding to the second component clothing.
- vi. This process is continued for all the components namely recreation, education, rent and miscellaneous.
- i. Shade/colour each sector with different shades/colours.

viii. Write legends to each component.

The pie diagram is presented in Fig 4.6.

The pie diagram is presented in Fig 4.6.



The central angle of the sector corresponding to the expenditure incurred on Education is 26°

Recreation cost is less than rent by 28°

Pictogram

Pictograms are diagrammatic representation of statistical data using pictures of resemblance. These are very useful in attracting attention. They are easily understood. For the purpose of propaganda, the pictorial presentations of facts are quite popular and they also find places in exhibitions. They are extensively used by the government organizations as well as by private institutions. If needed, scales can be fixed.

Despite its visual advantages, pictogram has limited application due to the usage of pictures resembling the data. It can express an approximate value than the given actual numerical value.






The following table gives the sugarcane production in tonnes per acre for various years.


Year	2013	2014	2015	2016	2017
Sugar Cane (in tonnes per acre)	10	13	9	15	18

Represent the above data by pictogram.

Solution :


The above data is represented by pictogram in the following manner:

2013	
2014	
2015	
2016	
2017	

 = 1 tonne

The Pictogram given below shows the number of persons who have traveled by train from Chennai to Rameshwaram on each day of a week.





 = 100 persons

From the Pictogram find:

- Number of travelers travelled during the week
- On which day there was a maximum rush in the train
- The difference between the maximum and minimum number of travelers.

Solution :

i. Here total number of  is 48, and each  represents 100 persons. Hence number of travelers travelled during the week is $48 \times 100 = 4800$.

ii. The maximum rush in the train is on Thursday.

iii. Maximum number of persons travelled on Thursday = 10

Hence the number of persons travelled on Thursday is $10 \times 100 = 1000$

Minimum number of persons travelled on Wednesday = 4

Hence the number of persons travelled on Wednesday is $4 \times 100 = 400$

Therefore difference between maximum and minimum number of travelers is $1000 - 400 = 600$ persons.

Types of Graphs

Graphical representation can be advantageous to bring out the statistical nature of the frequency distribution of quantitative variable, which may be discrete or continuous.

The most commonly used graphs are

1. **Histogram**
2. **Frequency Polygon**
3. **Frequency Curve**
4. **Cumulative Frequency Curves (Ogives)**

1. Histogram

A histogram is an attached bar chart or graph displaying the distribution of a frequency distribution in visual form. Take classes along the X-axis and the frequencies along the Y-axis.

Corresponding to each class interval, a vertical bar is drawn whose height is proportional to the class frequency.

Limitations:

We cannot construct a histogram for distribution with open-ended classes. The histogram is also quite misleading, if the distribution has unequal intervals.

Draw the histogram for the 50 students in a class whose heights (in cms) are given below.

Height	111 – 120	121 – 130	131 – 140	141 – 150	151 – 160	161 – 170
Number of students	4	11	15	9	8	3

Find the range, whose height of students are maximum.

Solution:

Since we are displaying the distribution of Height and Number of students in visual form, the histogram is drawn.

Step 1 : Heights are marked along the X-axis and labeled as “Height(in cms)”.

Step 2 : Number of students are marked along the Y-axis and labeled as “No. of students”.

Step 3 : Corresponding to each Heights, a vertical attached bar is drawn whose height is proportional to the number of students.

The Histogram is presented in Fig 4.7.

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

Height (in Cm)	No.of Students
110.5 - 120.5	4
120.5 - 130.5	11
130.5 - 140.5	15
140.5 - 150.5	9
150.5 - 160.5	8
160.5 - 170.5	3

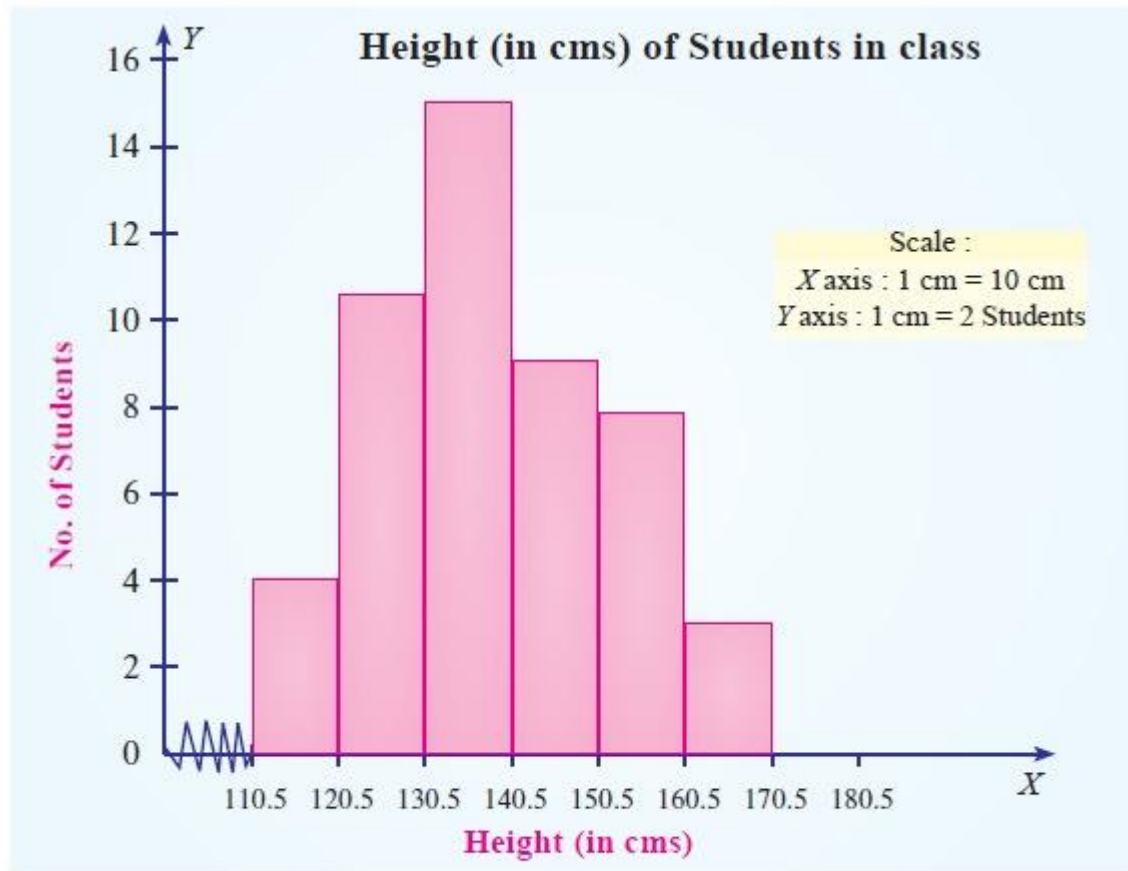


Fig 4.7 Histogram for heights of students in a class

The tallest bar shows that maximum number of students height are in the range 130.5 to 140.5 cm

The following table shows the time taken (in minutes) by 100 students to travel to school on a particular day

Time	0-5	5-10	10-15	15-20	20-25
No. of Students	5	25	40	17	13

Draw the histogram. Also find:

- The number of students who travel to school within 15 minutes.
- Number of students whose travelling time is more than 20 minutes.

Solution:

Since we are displaying the distribution of time taken (in minutes) by 100 students to travel to school on a particular day in visual form, the histogram is drawn.

Step 1 : Time taken are marked along the X-axis and labeled as “Time (in minutes)”.

Step 2 : Number of students are marked along the Y-axis and labeled as “No. of students”.

Step 3 : Corresponding to each time taken, a vertical attached bar is drawn whose height is proportional to the number of students.

The Histogram is presented in Fig 4.8.

i. $5+25+40=70$ students travel to school within 15 minutes

ii. 13 students travelling time is more than 20 minutes

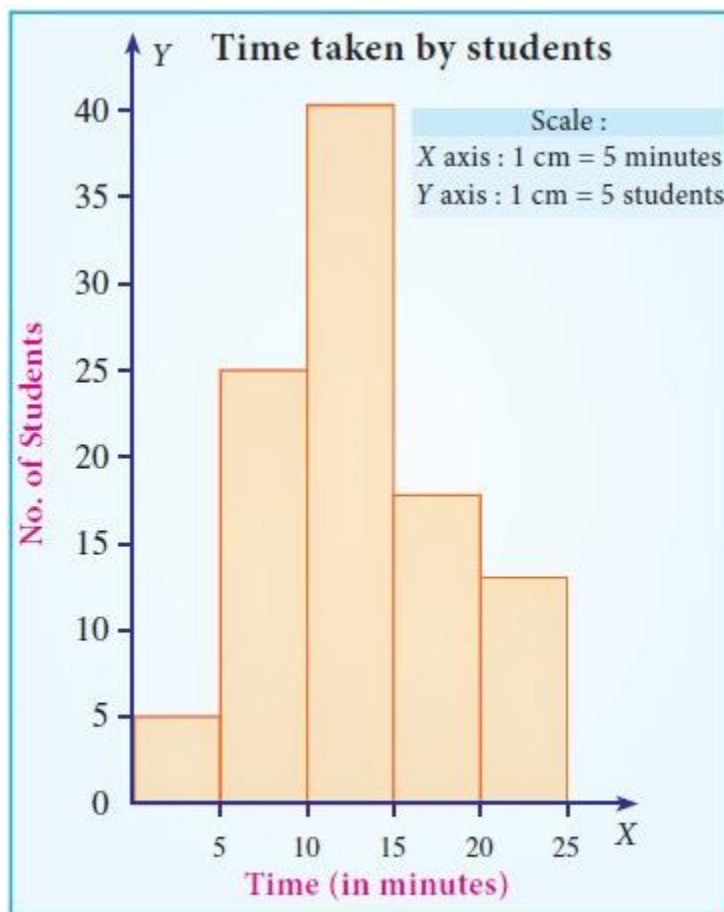


Fig 4.8 Histogram for time taken by students to travel to school

Draw a histogram for the following 100 persons whose daily wages (in `) are given below.

Daily wages	0 – 50	50 – 100	100 – 200	200 – 250	250 – 450	450 – 500
Number of persons	5	10	16	7	48	14

Also find:

- Number of persons who gets daily wages less than or equal to ` 200?
- Number of persons whose daily wages are more than ` 250?

Solution:

Since we are displaying the distribution of 100 persons whose daily wages in rupees in visual form, the histogram is drawn.

Step 1 : Daily wages are marked along the *X*-axis and labeled as “Daily Wages (in `)”.

Step 2 : Number of Persons are marked along the *Y*-axis and labeled as “No. of Persons”.

Step 3 : Corresponding to each daily wages, a vertical attached bar is drawn whose height is proportional to the number of persons.

The Histogram is presented in Fig 4.9.

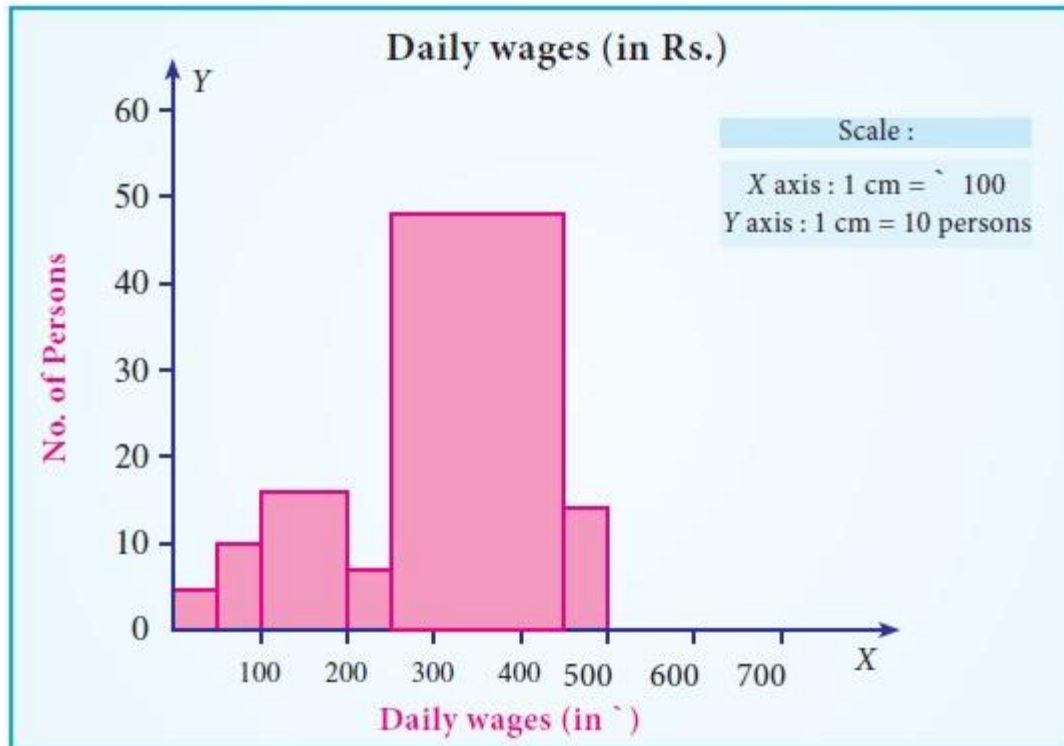


Fig 4.9 Histogram of daily wages (in ₹) for persons

- i. $5+10+16=31$ persons get daily wages less than or equal to ₹ 200.
- ii. $48+14=62$ persons get more than Rs 250.

2. Frequency Polygon

Frequency polygon is drawn after drawing histogram for a given frequency distribution. The area covered under the polygon is equal to the area of the histogram. Vertices of the polygon represent the class frequencies. Frequency polygon helps to determine the classes with higher frequencies. It displays the tendency of the data. The following procedure can be followed to draw frequency polygon:

- i. Mark the midpoints at the top of each vertical bar in the histogram representing the classes.
- ii. Connect the midpoints by line segments.

A firm reported that its Net Worth in the years 2011-2016 are as follows:

Year	2011 -2012	2012 – 2013	2013 – 2014	2014 – 2015	2015 - 2016
Net Worth (` in lakhs)	100	112	120	133	117

Draw the frequency polygon for the above data

Solution:

Since we are displaying the distribution of Net worth in the years 2011-2016, the Frequency polygon is drawn to determine the classes with higher frequencies. It displays the tendency of the data.

The following procedure can be followed to draw frequency polygon:

Step 1 : Year are marked along the X-axis and labeled as ‘Year’.

Step 2 : Net worth are marked along the Y-axis and labeled as ‘Net Worth (in lakhs of `)’.

Step 3 : Mark the midpoints at the top of each vertical bar in the histogram representing the year.

Step 4 : Connect the midpoints by line segments.

The Frequency polygon is presented in Fig 4.10.

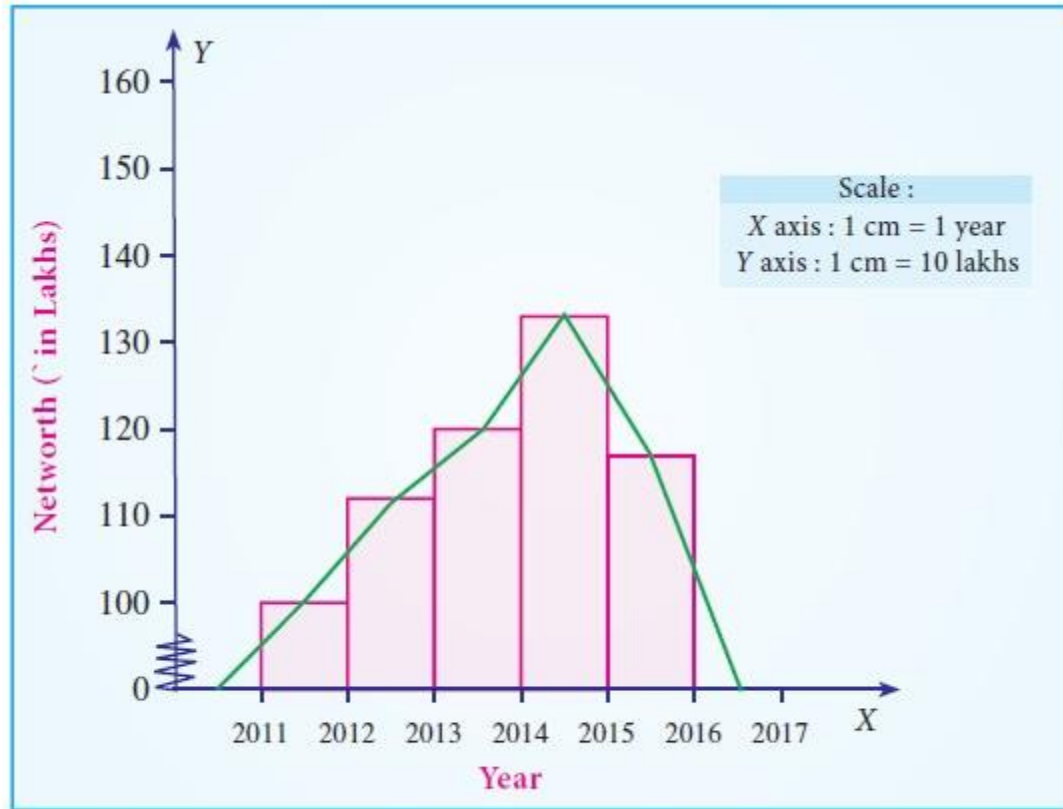


Fig 4.10 Frequency polygon for Net Worth in the years 2011-2016

3. Frequency Curve

Frequency curve is a smooth and free-hand curve drawn to represent a frequency distribution. Frequency curve is drawn by smoothing the vertices of the frequency polygon. Frequency curve provides better understanding about the properties of the data than frequency polygon and histogram.

The ages of group of pensioners are given in the table below. Draw the Frequency curve to the following data.

Age	65 - 70	70 - 75	75 - 80	80 - 85	85 - 90
No.of pensioners	38	45	24	10	8

Solution:

Since we are displaying the distribution of Age and Number of Pensioners, the Frequency curve is drawn, to provide better understanding about the age and number of pensioners than frequency polygon.

The following procedure can be followed to draw frequency curve:

Step 1 : Age are marked along the X-axis and labeled as 'Age'.

Step 2 : Number of pensioners are marked along the Y-axis and labeled as 'No. of Pensioners'.

Step3 : Mark the midpoints at the top of each vertical bar in the histogram representing the age.

Step 4 : Connect the midpoints by line segments by smoothing the vertices of the frequency polygon

The Frequency curve is presented in Fig 4.11.

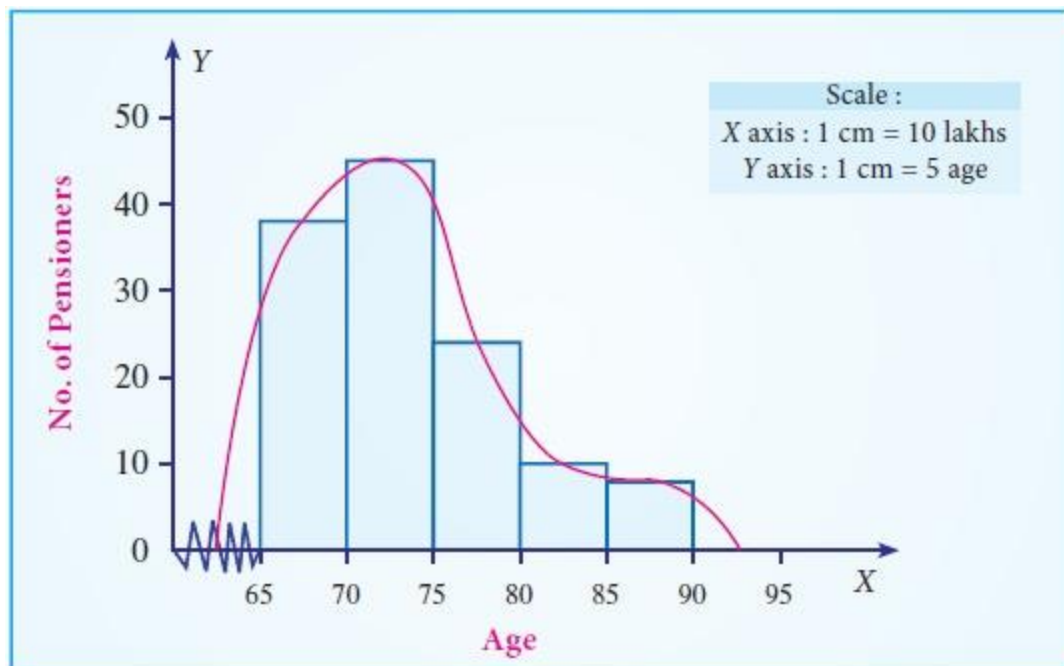


Fig 4.11 Frequency curve for Age and No. of pensioners

4. Cumulative frequency curve (Ogive)

Cumulative frequency curve (Ogive) is drawn to represent the cumulative frequency distribution. There are two types of Ogives such as 'less thanOgive curve' and 'more thanOgive curve'. To

draw these curves, we have to calculate the ‘less than’ cumulative frequencies and ‘more than’ cumulative frequencies. The following procedure can be followed to draw the ogive curves:

Less than Ogive: Less than cumulative frequency of each class is marked against the corresponding upper limit of the respective class. All the points are joined by a free-hand curve to draw **the less than ogive** curve.

More than Ogive: More than cumulative frequency of each class is marked against the corresponding lower limit of the respective class. All the points are joined by a free-hand curve to draw the **more than ogive** curve.

Both the curves can be drawn separately or in the same graph. If both the curves are drawn in the same graph, then the value of abscissa (x-coordinate) in the point of intersection is the median.

If the curves are drawn separately, median can be calculated as follows:

Draw a line perpendicular to Y-axis at $y=N/2$. Let it meet the Ogive at C. Then, draw a perpendicular line to X-axis from the point C. Let it meet the X-axis at M. The abscissa of M is the median of the data.

Draw the less than Ogive curve for the following data:

Daily Wages (in Rs.)	70- 80	80- 90	90-100	100-110	110-120	120-130	130-140	140-150
No. of workers	12	18	35	42	50	45	20	8

Also, find

i. The Median

ii. The number of workers whose daily wages are less than ` 125. [S](#)

Solution: Since we are displaying the distribution of Daily Wages and No. of workers, the Ogive curve is drawn, to provide better understanding about the wages and No. of workers.

The following procedure can be followed to draw Less than Ogive curve:

Step 1 : Daily wages are marked along the X-axis and labeled as “Wages(in `)”.

Step 2 : No. of Workers are marked along the Y-axis and labeled as “No. of workers”.

Step 3 : Find the less than cumulative frequency, by taking the upper class-limit of daily wages. The cumulative frequency corresponding to any upper class-limit of daily wages is the sum of all the frequencies less than the limit of daily wages.

Step 4 : The less than cumulative frequency of Number of workers are plotted as points against the daily wages (upper-limit). These points are joined to form less than ogive curve.

The Less than Ogive curve is presented in Fig 4.12.

Daily wages (less than)	No of workers
80	12
90	30
100	65
110	107
120	157
130	202
140	222
150	230

Daily Wages of Workers

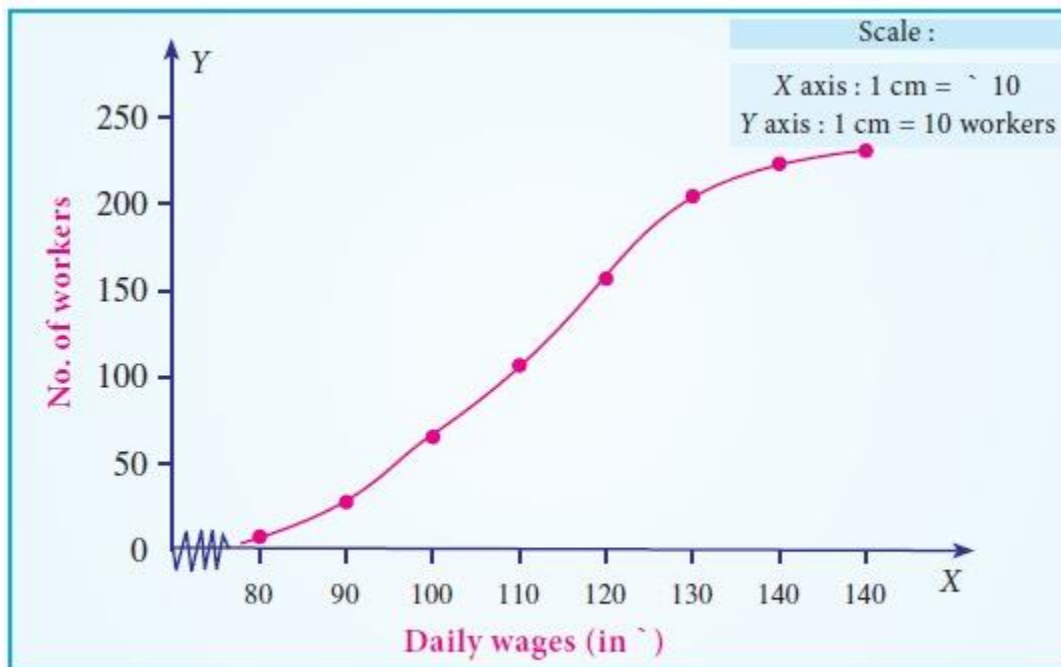


Fig 4.12 Less than Ogive curve for daily wages and number of workers

i. Median = ₹ 120

- ii. 183 workers get daily wages less than ` 125

The following table shows the marks obtained by 120 students of class IX in a cycle test-I . Draw the more than Ogive curve for the following data :

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No. of students	2	6	8	20	30	22	18	8	4	2

Also, find

- i. The Median
ii. The Number of students who get more than 75 marks.

Solution:

Since we are displaying the distribution marks and No. of students, the more than Ogive curve is drawn, to provide better understanding about the marks of the students and No. of students.

The following procedure can be followed to draw More than Ogive curve:

Step 1 : Marks of the students are marked along the X-axis and labeled as 'Marks'.

Step 2 : No. of students are marked along the Y-axis and labeled as 'No. of students'.

Step 3 : Find the more than cumulative frequency, by taking the lower class-limit of marks. The cumulative frequency corresponding to any lower class-limit of marks is the sum of all the frequencies above the limit of marks.

Step 4 : The more than cumulative frequency of number of students are plotted as points against the marks (lower-limit). These points are joined to form more than ogive curve.

The More than Ogive curve is presented in Fig 4.13.

Marks More than	No of Students
0	120
10	118

20	112
30	104
40	84
50	54
60	32
70	14
80	6
90	2

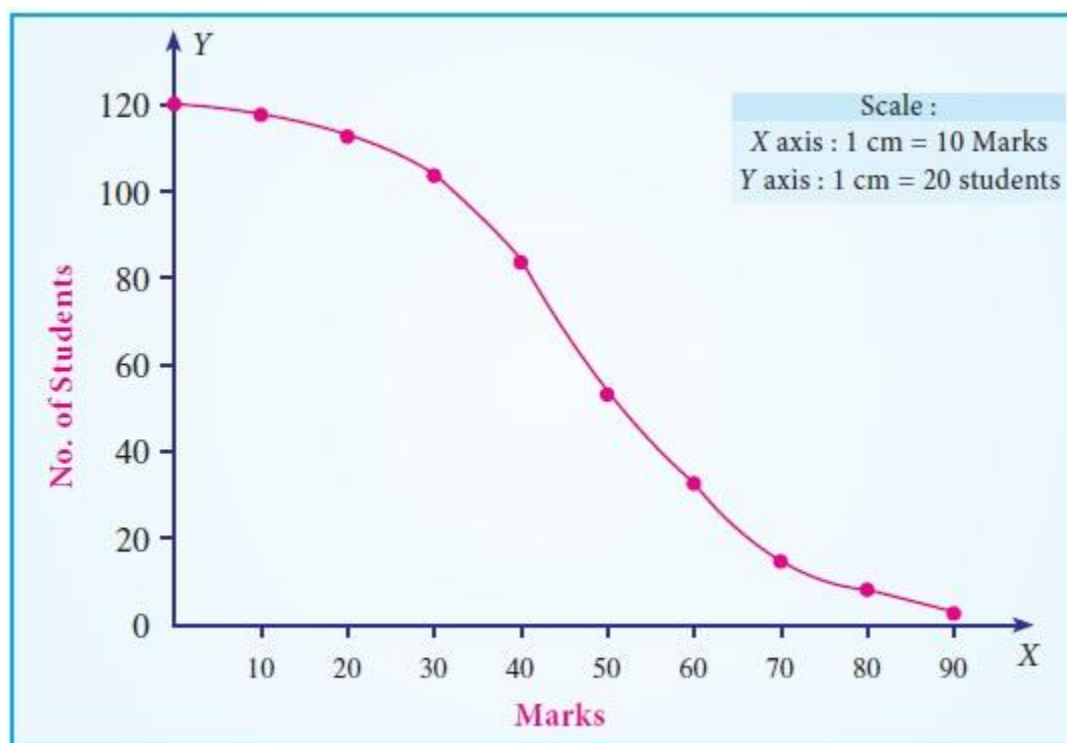


Fig 4.13 More than Ogive curve for Marks and No. of students

- Median = 42 students
- 7 students get more than 75 marks.

The yield of mangoes were recorded (in kg) are given below:

Graphically,

- i. find the number of trees which yield mangoes of less than 55 kg.
- ii. find the number of trees from which mangoes of more than 75 kg.
- iii. find the median.

Draw the Less than and More than Ogive curves. Also, find the median using the Ogive curves

Yield (in kg)	No. of trees
40 – 50	10
50 – 60	15
60 – 70	17
70 – 80	14
80 – 90	12
90 – 100	2
Total	70

Solution:

Since we are displaying the distribution of Yield and No. of trees, the Ogive curve is drawn, to provide better understanding about the Yield and No. of trees

The following procedure can be followed to draw Ogive curve:

Step 1 : Yield of mangoes are marked along the X-axis and labeled as ‘Yield (in Kg.)’.

Step 2 : No. of trees are marked along the Y-axis and labeled as ‘No. of trees’.

Step 3 : Find the less than cumulative frequency, by taking the upper class-limit of Yield of mangoes. The cumulative frequency corresponding to any upper class-limit of Mangoes is the sum of all the frequencies less than the limit of mangoes.

Step 4 : Find the more than cumulative frequency, by taking the lower class-limit of Yield of mangoes. The cumulative frequency corresponding to any lower class-limit of Mangoes is the sum of all the frequencies above the limit of mangoes.

Step 5 : The less than cumulative frequency of Number of trees are plotted as points against the yield of mangoes (upper-limit). These points are joined to form less than ogive curve.

Step 6 : The more than cumulative frequency of Number of trees are plotted as points against the yield of mangoes (lower-limit). These points are joined to form more than O give curve.

Less than Ogive		More than Ogive	
Yield less than	No. of trees	Yield greater than	No. of trees
50	10	40	70
60	25	50	60
70	42	60	45
80	56	70	28
90	68	80	14
100	70	90	2

The Ogive curve is presented in Fig 4.14.

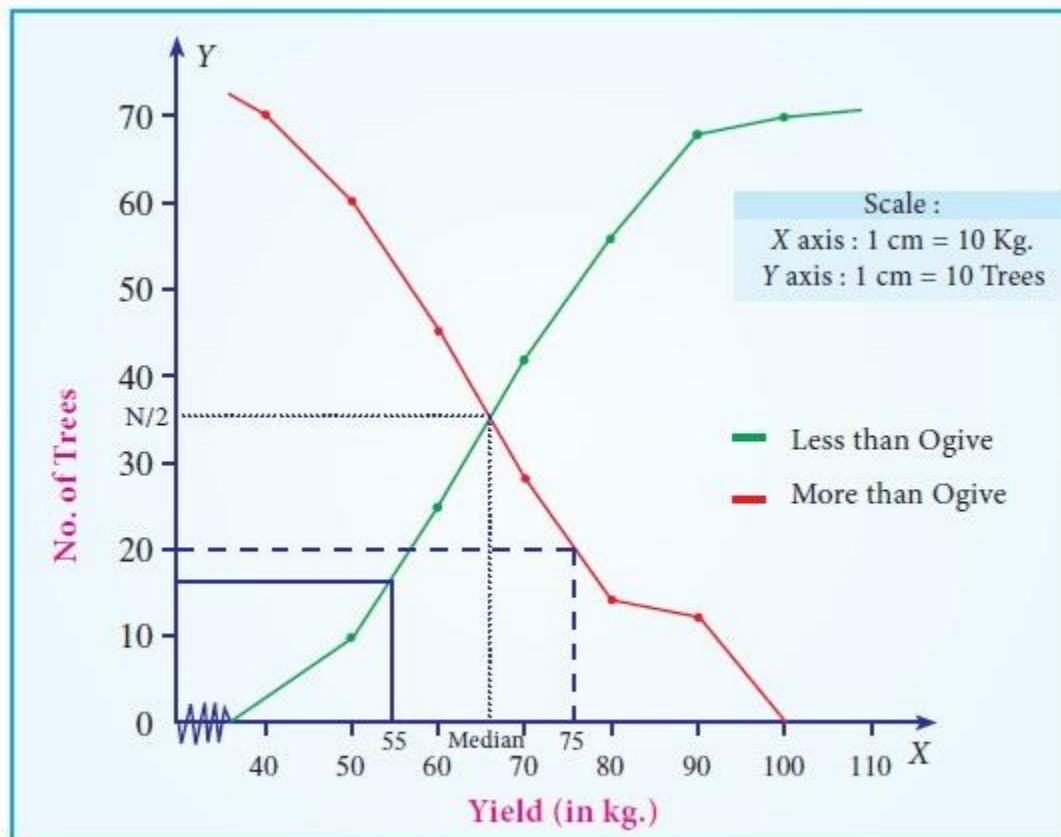


Fig 4.14 Ogive curve for Yield of mangoes and number of trees

- i. 16 trees yield less than 55 kg
- ii. 20 trees yield more than 75 kg
- iii. Median = 66 kg

Measures of Central Tendency

In practical situations one need to have a single value to represent each variable in the whole set of data. Because, the values of the variable are not equal, however there is a general tendency of such observations to cluster around a particular level. In this situation it may be preferable to characterize each group of observations by a single value such that all other values clustered around it. That is why such measure is called the measure of central tendency of that group. A measure of central tendency is a representative value of the entire group of data. It describes the characteristic of the entire mass of data. It reduces the complexity of data and makes them amenable for the application of mathematical techniques involved in analysis and interpretation of data.

Definition of Measures of Central Tendency

Various statisticians have defined the word average differently. Some of the important definitions are given below:

“Average is an attempt to find one single figure to describe whole of figure”

– *Clark and Sekkade*

“Average is a value which is typical or representative of a set of data”

– *Murray R. Speigal.*

“The average is sometimes described as number which is typical of the whole group”

– *Leabo.*

It is clear from the above definitions that average is a typical value of the entire data and is a measures of central tendency.

Characteristics for a good statistical average

The following properties should be possessed by an ideal average.

- ☐ It should be well defined so that a unique answer can be obtained.
- ☐ It should be easy to understand, calculate and interpret.
- ☐ It should be based on all the observations of the data.
- ☐ It should be amenable for further mathematical calculations.
- ☐ It should be least affected by the fluctuations of the sampling.
- ☐ It should not be unduly affected by the extreme values.

Arithmetic Mean

(a) To find A.M. for Raw dat

For a raw data, the arithmetic mean of a series of numbers is sum of all observations divided by the number of observations in the series. Thus if x_1, x_2, \dots, x_n represent the values of n observations, then arithmetic mean (A.M.) for n observations is: (direct method)

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

There are two methods for computing the A.M :

- (i) Direct method
- (ii) Short cut method.

The following data represent the number of books issued in a school library on selected from 7 different days 7, 9, 12, 15, 5, 4, 11 find the mean number of books.

Solution:

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ \bar{x} &= \frac{7+9+12+15+5+4+11}{7} \\ &= \frac{63}{7} = 9\end{aligned}$$

Hence the mean of the number of books is 9

Short-cut Method to find A.M.

Under this method an assumed mean or an arbitrary value (denoted by A) is used as the basis of calculation of deviations (d_i) from individual values. That is if $d_i = x_i - A$

Then

$$\bar{x} = A + \frac{\sum_{i=1}^n d_i}{n}$$

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

Solution:

Let us take the assumed mean, $A = 68$

x_i	$d_i = x_i - 68$
75	7
68	0
80	12
56	-12
92	24
Total	31

$$\begin{aligned}
 \bar{x} &= A + \frac{\sum_{i=1}^n d_i}{n} \\
 &= 68 + \frac{31}{5} \\
 &= 68 + 6.2 = 74.2
 \end{aligned}$$

The arithmetic mean of average marks is 74.2

(b) To find A.M. for Discrete Grouped data

If x_1, x_2, \dots, x_n are discrete values with the corresponding frequencies f_1, f_2, \dots, f_n . Then the mean for discrete grouped data is defined as (direct method)

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

In the short cut method the formula is modified as

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{N} \quad \text{where } d_i = x_i - A$$

A proof reads through 73 pages manuscript. The number of mistakes found on each of the pages are summarized in the table below Determine the mean number of mistakes found per page

No of mistakes	1	2	3	4	5	6	7
No of pages	5	9	12	17	14	10	6

Solution:

(i) Direct Method

x_i	f_i	$f_i x_i$
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
Total	N=73	299

$$\begin{aligned}
 \bar{x} &= \frac{\sum_{i=1}^n f_i x_i}{N} \\
 &= \frac{299}{73} \\
 &= 4.09
 \end{aligned}$$

The mean number of mistakes is 4.09

(ii) Short-cut Method

x_i	f_i	$d_i=x_i-A$	f_id_i
1	5	-3	-15
2	9	-2	-18
3	12	-1	-12
4	17	0	0
5	14	1	14
6	10	2	20
7	6	3	18
	$\Sigma f_i=73$		$\Sigma f_id_i=7$

$$\begin{aligned}
 \bar{x} &= A + \frac{\sum_{i=1}^n f_id_i}{N} \\
 &= 4 + \frac{7}{73} \\
 &= 4.09
 \end{aligned}$$

The mean number of mistakes = 4.09

(c) Mean for Continuous Grouped data:

For the computation of A.M for the continuous grouped data, we can use direct method or short cut method.

Direct Method:

The formula is

$$\bar{x} = \frac{\sum_{i=1}^n f_ix_i}{N}, \quad x_i \text{ is the midpoint of the class interval}$$

Short cut method

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{N} \times C$$

$$d = \frac{x_i - A}{c}$$

where A - any arbitrary value
 c - width of the class interval

x_i is the midpoint of the class interval.

The following the distribution of persons according to different income groups

Income (in ` 1000)	0 – 8	8 – 16	16 – 24	24 – 32	32 – 40	40 – 48
No of persons	8	7	16	24	15	7

Find the average income of the persons.

Solution :

Direct Method:

<i>Class</i>	f_i	x_i	$f_i x_i$
0-8	8	4	32
8 – 16	7	12	84
16-24	16	20	320
24-32	24	28	672
32-40	15	36	540
40-48	7	44	308
Total	N =77		1956

$$\begin{aligned}
 \bar{x} &= \frac{\sum_{i=1}^n f_i x_i}{N} \\
 &= \frac{1956}{77} \\
 &= 25.40
 \end{aligned}$$

Short cut method:

Class	f_i	x_i	$d_i = (x_i - A)/c$	$f_i d_i$
0 – 8	8	4	-3	-24
8 – 16	7	12	-2	-14
16 – 24	16	20	-1	-16
24 – 32	24	28 A	0	0
32 – 40	15	36	1	15
40 – 48	7	44	2	14
Total	N= 77			-25

$$\begin{aligned}\bar{x} &= A + \frac{\sum_{i=1}^n f_i d_i}{N} \times C \\ &= 28 + \frac{-25}{77} \times 8 = 25.40\end{aligned}$$

Merits

- ☐ It is easy to compute and has a unique value.
- ☐ It is based on all the observations.
- ☐ It is well defined.
- ☐ It is least affected by sampling fluctuations.
- ☐ It can be used for further statistical analysis.

Limitations

- ☐ The mean is unduly affected by the extreme items (outliers).
- ☐ It cannot be determined for the qualitative data such as beauty, honesty etc
- ☐ It cannot be located by observations on the graphic method.

When to use?

Arithmetic mean is a best representative of the data if the data set is homogeneous. On the other hand if the data set is heterogeneous the result may be misleading and may not represent the data.

Weighted Arithmetic Mean

The arithmetic mean, as discussed earlier, gives equal importance (or weights) to each observation in the data set. However, there are situations in which values of individual observations in the data set are not of equal importance. Under these circumstances, we may attach, a weight, as an indicator of their importance to each observation value.

Definition

Let x_1, x_2, \dots, x_n be the set of n values having weights w_1, w_2, \dots, w_n respectively, then the weighted mean is,

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Uses of weighted arithmetic mean

Weighted arithmetic mean is used in:

- ☐ The construction of index numbers.
- ☐ Comparison of results of two or more groups where number of items in the groups differs.
- ☐ Computation of standardized death and birth rates.

The weights assigned to different components in an examination or Component Weightage
Marks scored

Component	Weightage	Marks scored
Theory	4	60
Practical	3	80
Assignment	1	90
Project	2	75
	10	

Calculate the weighted average score of the student who scored marks as given in the table

Solution:

Component	Marks scored (x_i)	Weightage (w_i)	$w_i x_i$
Theory	60	4	240
Practical	80	3	240
Assignment	90	1	90
Project	75	2	150
Total		10	720

$$\begin{aligned}
 \text{Weighted average, } \bar{x} &= \frac{\sum w_i x_i}{\sum w_i} \\
 &= 720/10 \\
 &= 72
 \end{aligned}$$

Combined Mean:

Let \bar{x}_1 and \bar{x}_2 are the arithmetic mean of two groups (having the same unit of measurement of a variable), based on n_1 and n_2 observations respectively. Then the combined mean can be calculated using

$$\text{Combined Mean} = \bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Remark : The above result can be extended to any number of groups.

A class consists of 4 boys and 3 girls. The average marks obtained by the boys and girls are 20 and 30 respectively. Find the class average.

Solution:

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

$$\begin{aligned} \text{Combined Mean} = \bar{x}_{12} &= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \\ &= \left[\frac{4 \times 20 + 3 \times 30}{4 + 3} \right] \\ &= \left[\frac{80 + 90}{7} = \frac{170}{7} \right] = 24.3 \end{aligned}$$

Median

Median is the value of the variable which divides the whole set of data into two equal parts. It is the value such that in a set of observations, 50% observations are above and 50% observations are below it. Hence the median is a positional average.

(a) Median for Ungrouped or Raw data:

In this case, the data is arranged in either ascending or descending order of magnitude.

(i) If the number of observations n is an odd number, then the median is represented by the numerical value of x , corresponds to the positioning point of $n+1 / 2$ in ordered observations. That is,

Median = value of $(n+1 / 2)^{th}$ observation in the data array

If the number of observations n is an even number, then the median is defined as the arithmetic mean of the middle values in the array That is,

$$\text{Median} = \frac{\text{value of } \left(\frac{n}{2}\right)^{th} \text{ observation} + \text{value of } \left(\frac{n}{2} + 1\right)^{th} \text{ observation}}{2}$$

Example

The number of rooms in the seven five stars hotel in Chennai city is 71, 30, 61, 59, 31, 40 and 29. Find the median number of rooms

Solution:

Arrange the data in ascending order 29, 30, 31, 40, 59, 61, 71

$n = 7$ (odd)

Median = $7+1 / 2 = 4^{th}$ positional value

Median = 40 rooms

The export of agricultural product in million dollars from a country during eight quarters in 1974 and 1975 was recorded as 29.7, 16.6, 2.3, 14.1, 36.6, 18.7, 3.5, 21.3

Find the median of the given set of values

Solution:

We arrange the data in descending order

36.6, 29.7, 21.3, 18.7, 16.6, 14.1, 3.5, 2.3

$$n = 8 \text{ (even)}$$

$$\begin{aligned}\text{Median} &= \frac{4^{\text{th}} \text{ item} + 5^{\text{th}} \text{ item}}{2} \\ &= \frac{18.7 + 16.6}{2} \\ &= 17.65 \text{ million dollars}\end{aligned}$$

Cumulative Frequency

In a grouped distribution, values are associated with frequencies. The cumulative frequencies are calculated to know the total number of items above or below a certain limit. This is obtained by adding the frequencies successively up to the required level. These cumulative frequencies are useful to calculate median, quartiles, deciles and percentiles.

(b) Median for Discrete grouped data

We can find median using following steps

- i. Calculate the cumulative frequencies
- ii. Find $(N+1)/2$, where $N = \sum f = \text{total frequencies}$
- iii. Identify the cumulative frequency just greater than $(N+1)/2$
- iv. The value of x corresponding to that cumulative frequency is the $(N+1)/2$ median.

The following data are the weights of students in a class. Find the median weights of the students

Weight(kg)	10	20	30	40	50	60	70
Number of Students	4	7	12	15	13	5	4

Solution:

Weight (kg) x	Frequency f	Cumulative Frequency $c.f$
10	4	4
20	7	11
30	12	23
40	15	38
50	13	51
60	5	56
70	4	60
Total	N = 60	

Here, $N = \sum f = 60$

$$\frac{N+1}{2} = 30.5$$

The cumulative frequency greater than 30.5 is 38. The value of x corresponding to 38 is 40. The median weight of the students is 40 kgs

(c) Median for Continuous grouped data

In this case, the data is given in the form of a frequency table with class-interval etc., The following formula is used to calculate the median.

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

Where

l = Lower limit of the median class

N = Total Numbers of frequencies

f = Frequency of the median class

m = Cumulative frequency of the class preceding the median class

c = the class interval of the median class.

From the formula, it is clear that one has to find the median class first. Median class is, that class which correspond to the cumulative frequency just greater than $N/2$.

The following data attained from a garden records of certain period Calculate the median weight of the apple

Weight in grams	410 – 420	420 – 430	430 – 440	440 – 450	450 – 460	460 – 470	470 – 480
Number of apples	14	20	42	54	45	18	7

Solution:

Weight in grams	Number of apples	Cumulative Frequency
410 – 420	14	14
420 – 430	20	34
430 – 440	42	76
440 – 450	54	130
450 – 460	45	175
460 – 470	18	193
470 – 480	7	200
Total	N = 200	

$$\frac{N}{2} = \frac{200}{2} = 100.$$

Median class is 440 – 450

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

$$l = 440, \quad \frac{N}{2} = 100, \quad m = 76, \quad f = 54, \quad c = 10$$

$$\begin{aligned} \text{Median} &= 440 + \frac{100 - 76}{54} \times 10 \\ &= 440 + \frac{24}{54} \times 10 = 440 + 4.44 = 444.44 \end{aligned}$$

The median weight of the apple is 444.44 grams

The following table shows age distribution of persons in a particular region:

Age (years)	No. of persons (in thousands)
Below 10	2
Below 20	5
Below 30	9
Below 40	12
Below 50	14
Below 60	15
Below 70	15.5
Below 80	15.6

Find the median age.

Solution:

We are given upper limit and less than cumulative frequencies. First find the class-intervals and the frequencies. Since the values are increasing by 10, hence the width of the class interval is equal to 10.

Age groups	No. of persons (in thousands) f	cf
0 – 10	2	2
10 – 20	3	5
20 – 30	4	9
30 – 40	3	12
40 – 50	2	14
50 – 60	1	15
60 – 70	0.5	15.5
70 – 80	0.1	15.6
Total	N = 15.6	

$$\left(\frac{N}{2}\right) = \frac{15.6}{2} = 7.8$$

Median lies in the 20 – 30 age group

$$\begin{aligned}\text{Median} &= l + \frac{\frac{N}{2} - m}{f} \times c \\ &= 20 + \frac{7.8 - 5}{4} \times 10\end{aligned}$$

Median = 27 years

The following is the marks obtained by 140 students in a college. Find the median marks

Marks	Number of students
10-19	7
20-29	15
30-39	18
40-49	25
50-59	30
60-69	20
70-79	16
80-89	7
90-99	2

Solution:

Class boundaries	f	Cf
9.5 -19.5	7	7
19.5-29.5	15	22
29.5- 39.5	18	40
39.5-49.5	25	65
49.5-59.5	30	95
59.5-69.5	20	115
69.5-79.5	16	131
79.5-89.5	7	138
89.5-99.5	2	140
Total	N =140	

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

$$\frac{N}{2} = \frac{140}{2} = 70$$

Here $l = 49.5$, $f = 30$, $m = 65$, $c = 10$

$$\begin{aligned} \text{Median} &= 49.5 + \left(\frac{70 - 65}{30} \right) \times 10 \\ &= 49.5 + 1.67 \\ &= 51.17 \end{aligned}$$

Graphical method for Location of median

Median can be located with the help of the cumulative frequency curve or ‘ogive’.

The procedure for locating median in a grouped data is as follows:

Step 1 : The class intervals, are represented on the horizontal axis (x-axis)

Step 2 : The cumulative frequency corresponding to different classes is calculated. These cumulative frequencies are plotted on the vertical axis (y-axis) against the upper limit of the respective class interval

Step 3 : The curve obtained by joining the points by means of freehand is called the ‘*less than ogive*’.

Step 4 : A horizontal straight line is drawn from the value $N/2$ and $N+1 / 2$ on the y-axis parallel to x- axis to meet the ogive. (depending on N is odd or even)

Step 5 : From the point of intersection, draw a line, perpendicular to the horizontal axis which meet the x axis at m say.

Step 6 : The value m at x axis gives the value of the median.

Remarks:

- (i) Similarly 'more than' ogives, can be drawn by plotting more than cumulative frequencies against lower limit of the class. A horizontal straight line is drawn from the value $\frac{N}{2}$ or $\frac{N+1}{2}$ on the y -axis parallel to x -axis to meet the ogive. A line is drawn perpendicular to x -axis meets the point at m , say, the X coordinate of m gives the value of the median.

(depending on N is odd or even)

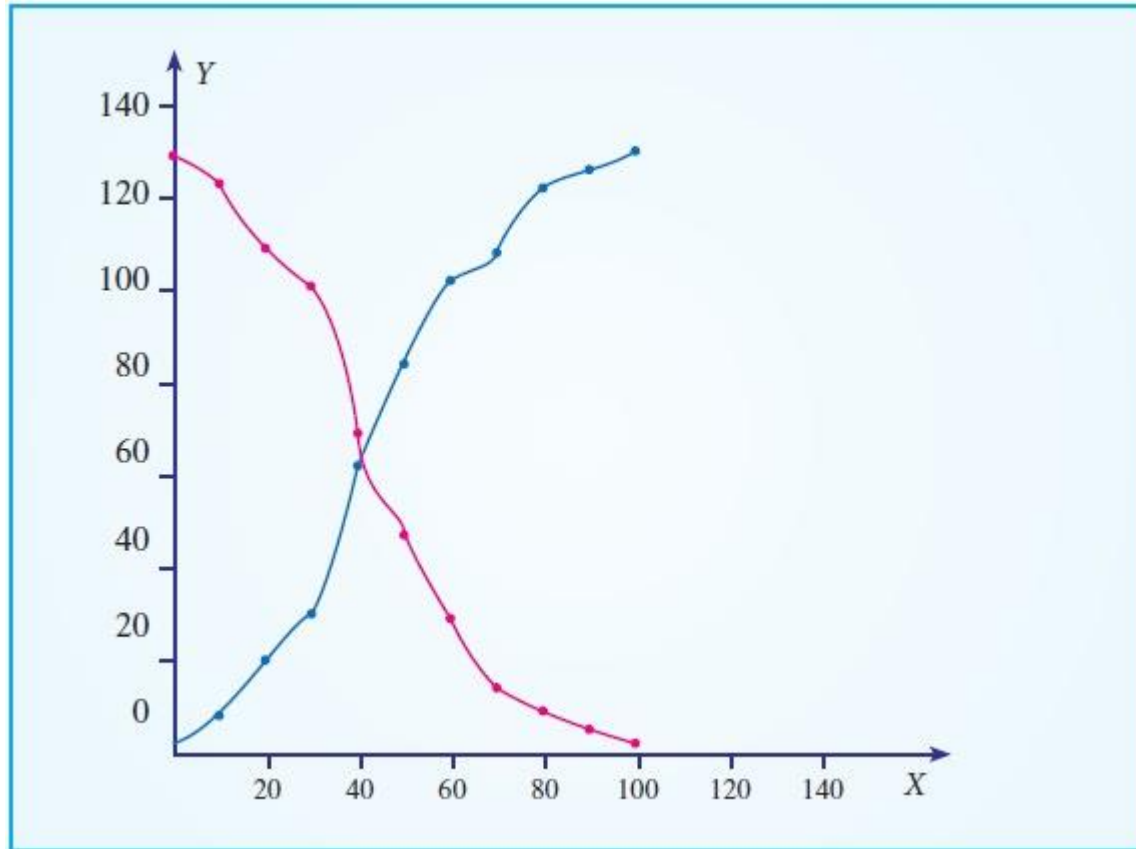
- (ii) When the two ogive curves are drawn on the same graph, a line is drawn perpendicular to x -axis from the point of intersection, meets the point at m , say. The x coordinate m gives the value of the median.

Draw ogive curves for the following frequency distribution and determine the median.

Age groups	No. of people
0 – 10	6
10 – 20	12
20 – 30	10
30 – 40	32
40 – 50	22
50 – 60	18
60 – 70	15
70 – 80	5
80 – 90	4
90 – 100	3

Solution:

Class boundary	Cumulative Frequency	
	Less than	More than
0	0	127
10	6	121
20	18	109
30	28	99
40	60	67
50	82	45
60	100	27
70	115	12
80	120	7
90	124	3
100	127	0



The median value from the graph is 42

Merits

- ☐ It is easy to compute. It can be calculated by mere inspection and by the graphical method
- ☐ It is not affected by extreme values.
- ☐ It can be easily located even if the class intervals in the series are unequal

Limitations

- ☐ It is not amenable to further algebraic treatment
- ☐ It is a positional average and is based on the middle item
- ☐ It does not take into account the actual values of the items in the series

Mode

According to Croxton and Cowden, ‘The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated.

In a busy road, where we take a survey on the vehicle - traffic on the road at a place at a particular period of time, we observe the number of two wheelers is more than cars, buses and other vehicles. Because of the higher frequency, we say that the modal value of this survey is ‘two wheelers’

Mode is defined as the value which occurs most frequently in a data set. The mode obtained may be two or more in frequency distribution.



Computation of mode:

(a) For Ungrouped or Raw Data:

The mode is defined as the value which occurs frequently in a data set

The following are the marks scored by 20 students in the class. Find the mode 90, 70, 50, 30, 40, 86, 65, 73, 68, 90, 90, 10, 73, 25, 35, 88, 67, 80, 74, 46

Solution:

Since the marks 90 occurs the maximum number of times, three times compared with the other numbers, mode is 90.

A doctor who checked 9 patients’ sugar level is given below. Find the mode value of the sugar levels. 80, 112, 110, 115, 124, 130, 100, 90, 150, 180

Solution:

Since each values occurs only once, there is no mode.

Compute mode value for the following observations.

2, 7, 10, 12, 10, 19, 2, 11, 3, 12

Solution:

Here, the observations 10 and 12 occurs twice in the data set, the modes are 10 and 12.

For discrete frequency distribution, mode is the value of the variable corresponding to the maximum frequency.

Example

Calculate the mode from the following data

Days of Confinement	6	7	8	9	10
Number of patients	4	6	7	5	3

Solution:

Here, 7 is the maximum frequency, hence the value of x corresponding to 7 is 8.

Therefore 8 is the mode.

(b) Mode for Continuous data:

The mode or modal value of the distribution is that value of the variate for which the frequency is maximum. It is the value around which the items or observations tend to be most heavily concentrated. The mode is computed by the formula.

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

Modal class is the class which has maximum frequency.

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

c = width of the class limits

Remarks

- (i) If $(2f_1 - f_0 - f_2)$ comes out to be zero, then mode is obtained by the following formula taking absolute differences $M_0 = l + \left(\frac{f_1 - f_0}{|f_1 - f_0| + |f_1 - f_2|} \times C \right)$
- (ii) If mode lies in the first class interval, then f_0 is taken as zero.
- (iii) The computation of mode poses problem when the modal value lies in the open-ended class.

The following data relates to the daily income of families in an urban area. Find the modal income of the families.

Income (`)	0-100	100-200	200-300	300-400	400-500	500-600	600-700
No.of persons	5	7	12	18	16	10	5

Solution:

Income (`)	No.of persons (f)
0-100	5
100-200	7
200-300	12 f_0
300-400	18 f_1
400-500	16 f_2
500-600	10
600-700	5

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

The highest frequency is 18, the modal class is 300-400

Here, $l = 300$, $f_0 = 12$, $f_1 = 18$, $f_2 = 16$,

$$\begin{aligned}
 \text{Mode} &= 300 + \frac{18 - 12}{2 \times 18 - 12 - 16} \times 100 \\
 &= 300 + \frac{6}{36 - 28} \times 100 \\
 &= 300 + \frac{6}{8} \times 100 \\
 &= 300 + \frac{600}{8} = 300 + 75 = 375
 \end{aligned}$$

The modal income of the families is 375.

Determination of Modal class:

For a frequency distribution modal class corresponds to the class with maximum frequency. But in any one of the following cases that is not easily possible.

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

Steps for preparing Analysis table:

We prepare a grouping table with 6 columns

- i. In column I, we write down the given frequencies.
- ii. Column II is obtained by combining the frequencies two by two.
- iii. Leave the 1st frequency and combine the remaining frequencies two by two and write in column III
- iv. Column IV is obtained by combining the frequencies three by three.
- v. Leave the 1st frequency and combine the remaining frequencies three by three and write in column V
- vi. Leave the 1st and 2nd frequencies and combine the remaining frequencies three by three and write in column VI

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

Calculate mode for the following frequency distribution:

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

Solution:

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

Analysis Table:

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

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

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

Here, $l = 20$, $f_0 = 16$, $f_1 = 17$, $f_2 = 15$

$$= 20 + \frac{17 - 16}{2 \times 17 - 16 - 15} \times C$$

$$= 20 + \frac{1}{34 - 31} \times 5$$

$$= 20 + \frac{5}{3} = 20 + 1.67 = 21.67$$

$$\text{Mode} = 21.67$$

(d) Graphical Location of Mode

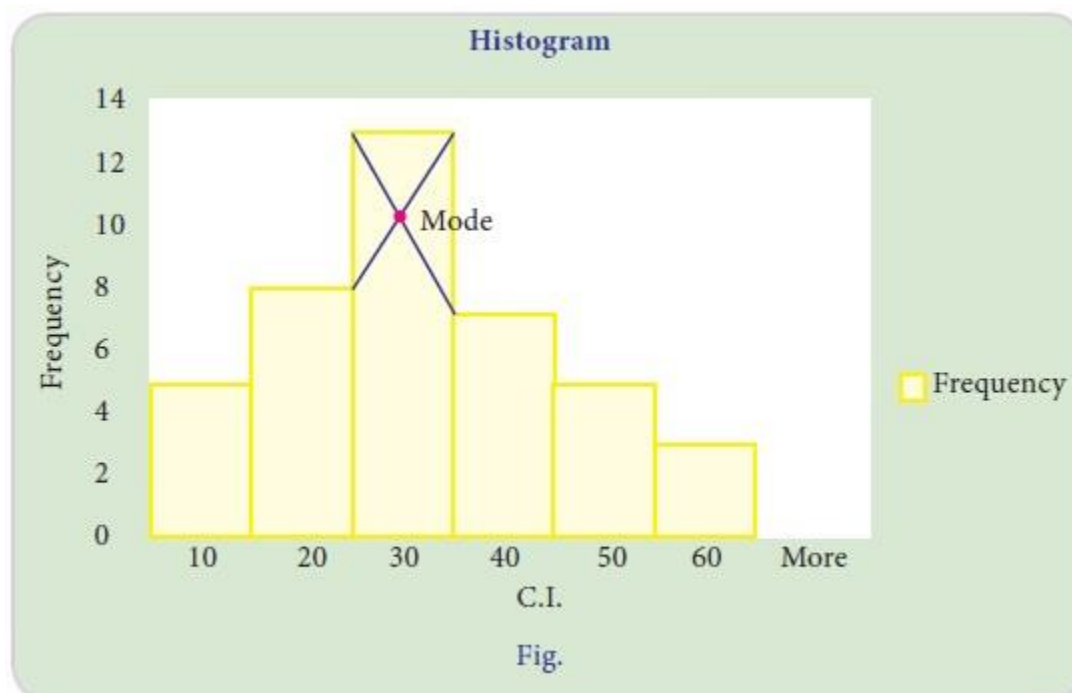
The following are the steps to locate mode by graph

- Draw a histogram of the given distribution.
- Join the rectangle corner of the highest rectangle (modal class rectangle) by a straight line to the top right corner of the preceding rectangle. Similarly the top left corner of the highest rectangle is joined to the top left corner of the rectangle on the right.
- From the point of intersection of these two diagonal lines, draw a perpendicular line to the x –axis which meets at M.
- The value of x coordinate of M is the mode.

Locate the modal value graphically for the following frequency distribution

Class Interval	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
Frequency	5	8	12	7	5	3

Solution:



Merits of Mode:

- ☐ It is comparatively easy to understand.
- ☐ It can be found graphically.
- ☐ It is easy to locate in some cases by inspection.
- ☐ It is not affected by extreme values.
- ☐ It is the simplest descriptive measure of average.

Demerits of Mode:

- ☐ It is not suitable for further mathematical treatment.
- ☐ It is an unstable measure as it is affected more by sampling fluctuations.
- ☐ Mode for the series with unequal class intervals cannot be calculated.
- ☐ In a bimodal distribution, there are two modal classes and it is difficult to determine the values of the mode.



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

QUESTION BANK

UNIT I PART-A

1. What are the common measures of central tendency?
2. Define mean, median and mode.
3. The weekly wages of 10 workers are 25,30,32,40,47,48,50,55,65. Find the average wage per week.
4. State the empirical relationship between the averages.
5. This is about marks of 60 students in a class. Calculate the standard deviation of marks.

Marks	10	20	30	40	50	50
No. of students	8	12	20	10	7	3

6. Find the range and coefficient of range for the following data
25,36,41,3,22,46,24,2,40,36,28,31,45,2,34
7. Find the S.D of the following set of observations 26,24,29,22,30,19,24,28,26,30
8. Define the term kurtosis.
9. Find the mode for the given observations 52,75,40,70,43,40,65,35,48.
10. Write the empirical relation between mean, median and mode.

PART B

1. Interpret different types of Graphs in detail with examples.
2. Explain different type of classification with Explain.
3. Explain different types of diagrams in detail
4. Construct Ogive curves for the following frequency distribution and determine median.

C.I: 0-10 10-20 20-30 30-40 40-50 50-60 60-70 70-80 80-90 90-100

F : 6 12 10 32 22 18 15 5 4 3

5. This is about marks of 60 students in a class. Calculate the standard deviation of marks.

Marks	10	20	30	40	50	50
No. of students	8	12	20	10	7	3

6. Calculate the mean, median and mode for the following data

Marks	8	10	12	15	18
No of students	5	7	12	6	10

7. Obtain mean, median and mode for the following frequency distribution

Class interval	20 - 30	30- 40	40 – 50	50 - 60	60 – 70	70- 80	80- 90	90- 100
frequency	3	8	9	15	20	13	8	4

8. Calculate the mean and median from the following distribution

C. I	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	6	12	22	48	56	32	18	6

9. This is about marks of 60 students in a class. Calculate the standard deviation of marks.

Marks	10	20	30	40	50	50
No. of students	8	12	20	10	7	3

10. Enumerate various measures of central tendencies in statistics.

UNIT – II – Correlation and Regression & Curve Fitting – SMTA1207

I. Introduction

CORRELATION AND REGRESSION & CURVE FITTING

CORRELATION

It refers the combination of two words ‘co’ (together) and relation (connection) between two quantities. Correlation is the statistical tool to measure the degree to which two variables are linearly related to each other.

i.e., To study the relationship between two variables.

If the quantities (X,Y) vary in such way that change in one variable corresponds to change in other variable, then the variable X and Y are correlated.

Example: Price of crude oil and stock price of an oil producing company.

Price of commodity and amount of demand.

Years of Experience and Salary of Employees

Dividend and Premium of Shares

Population and National Income etc.

Types of Correlation:

- (i) Positive Correlation

If for an increase in the value of one variable there is also an increase in the value of other variable and vice versa. (Same Direction)

***Examples:* The more time spend running on a treadmill, the more calories burn out.**

Time spend on Marketing and Customers etc.

Temperature and Ice cream sales.

(i) Negative Correlation:

If for an increase in the value of one variable there is a decrease in the value of the other variable and vice versa.(Opposite Direction)

***Examples:* Quantity of a commodity demanded and its price are**

Negatively correlated.

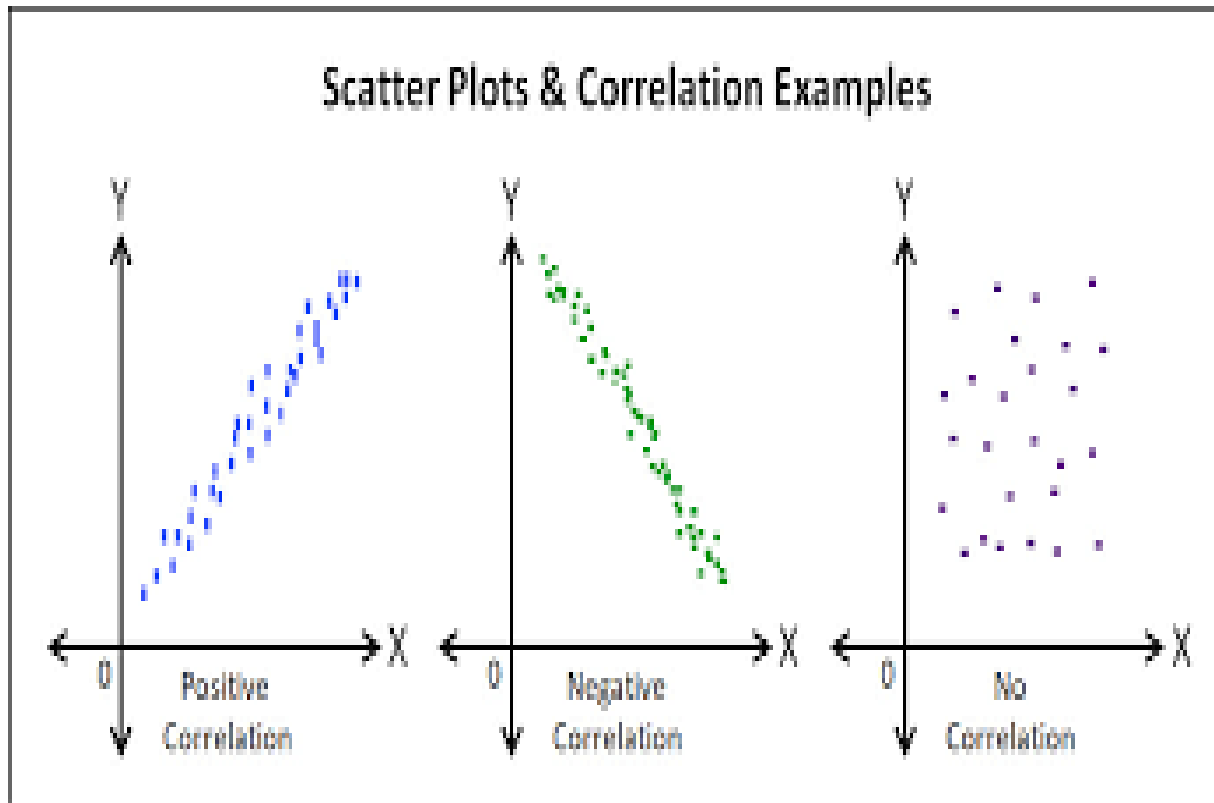
Tax and dividend.

(ii) No correlation:

If the change in the value of one variable has no connection with the change in the value of other variables.

***Examples:* Shoe size and Salary**

Weight of person and Colour of his hair.



Correlation coefficient: A Correlation coefficient is a numerical measure of some type of correlation. It is a statistical measure of the strength of the relationship between the two variables.

Properties of correlation coefficient

The coefficient of correlation lies between -1 and +1.

When r is positive, the variables x and y increase or decrease together.

If $r = +1$ then there is a perfect positive correlation.

When r is negative the variables x and y move in opposite direction.

If $r = -1$ then there is a perfect negative correlation.

If $r = 0$ then the variables are uncorrelated.

Problems

Q1: Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y).

X:	65	66	67	67	68	69	70	72
Y:	67	68	65	68	72	72	69	71

Solution:

Formula (Karl – Pearson’s Coefficient of Correlation)

$$n\sum XY - (\sum X)(\sum Y)$$

$$r_{XY} = \frac{\quad}{\quad}$$

$$\frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

X	Y	X²	Y²	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
$\sum X=544$	$\sum Y=552$	$\sum X^2=37028$	$\sum Y^2=38132$	$\sum XY=37560$

Here n=8.

Substituting above values in the formula,

We get

$$\frac{8(37560) - (544)(552)}{\quad}$$

$$r_{xy} = \frac{\quad}{\quad}$$

$$\frac{\text{SQRT}[8(37028)-(544)^2]}{\text{SQRT}[8(38132)-(552)^2]}$$

$$= 0.603$$

There is a positive correlation between x and y.

Q2. A computer while calculating the correlation coefficient between x and y from 25 pairs of observations, obtained the following

$$n=25, \sum x=125, \sum x^2=650, \sum y=100, \sum y^2=460, \sum xy=508.$$

It was however, later discovered at the time of checking that they had copied down two pairs as (6,14) and (8,6) while the correct values were (8,12) and (6,8). Obtain the correct value of the correlation coefficient.

Solution:

The correct values are $\sum x=125-6-8+8+6=125$

$$\sum y=100-14-6+12+8=100$$

$$\sum x^2=650-6^2-8^2+8^2+6^2=650$$

$$\sum y^2=460-14^2-6^2+12^2+8^2=436$$

$$\sum xy=508-(6 \times 14)-(8 \times 6)+(8 \times 12)+(6 \times 8)=520$$

Therefore,

The correct value of correlation coefficient

$$r_{XY} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

$$r_{XY} = \frac{(25)(520) - (125)(100)}{\sqrt{[(25)(650) - (125)^2][(25)(436) - (100)^2]}}$$

$$= 0.667.$$

RANK CORRELATION

It is a Qualitative assessment measurement of analyzing data arranged in order of merit in possession of two characteristics A and B.

Examples: Honesty, Beauty, Intelligence etc.,

In general the assumption that the values of variables are exactly measurable. In some situations, it may not be possible to give precise values for the variables. In such cases we can use another measure of correlation coefficient called rank correlation.

Let (x_i, y_i) $i = 1, 2, 3, \dots, n$ be the ranks of n individuals in the group for two characteristics A and B respectively. The correlation coefficient between the x_i, y_i is called the rank correlation.

Spearman's Rank Correlation coefficient

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

where $d_i = x_i - y_i$ and n is the number of pairs of observations.

Types:

1. When ranks are given
2. When the ranks are not given
3. When equal ranks are given.

PROBLEMS

Q1. When ranks are given:

The following are the ranks obtained by 10 students in statistics and mathematics. To what extent is knowledge of students in statistics related to knowledge in mathematics?

Rank of Stats	: 1	2	3	4	5	6	7	8	9	10
Rank of Maths	:2	4	1	5	3	9	7	10	6	8

Solution:

Rank in Statistics(R₁)	Rank in Mathematics (R₂)	d=x-y	d²
1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4

			$\Sigma d^2=40$

$$\rho_{xy} = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 40}{10(100-1)} = 0.76.$$

Q2. When ranks are not given:

Calculate Spearman's rank correlation for the following data.

X: 53 98 95 81 75 71 59 55

Y: 47 25 32 37 30 40 39 45

Solution:

X	Y	Rank X (R₁)	Rank Y (R₂)	d=(R₁-R₂)	d²
53	47	8	1	7	49
98	25	1	8	-7	49
95	32	2	6	-4	16
81	37	3	5	-2	4
75	30	4	7	-3	9
71	40	5	3	2	4
59	39	6	4	2	4
55	45	7	2	5	25

$$\sum d_i^2 = 160$$

$$\rho_{XY} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 160}{8(64 - 1)} = -0.9048.$$

There is very high negative correlation between X and Y.

Equal Ranks:

Q3.Find the rank correlation coefficient for the following data

x	92	89	87	86	86	77	71	63	53	50
y	86	83	91	77	68	85	52	82	37	57

Solution:

Let R_1 and R_2 denote the ranks in x and y respectively.

x	Y	R_1	R_2	$d=R_1-R_2$	d^2
92	86	1	2	-1	1
89	83	2	4	-2	4
87	91	3	1	2	4
86	77	4.5	6	-1.5	2.25
86	68	4.5	7	-2.5	6.25
77	85	6	3	3	9.00
71	52	7	9	-2	4.00
63	82	8	5	3	9.00
53	37	9	10	-1	1.00
50	57	10	8	2	4.00
					$\sum d_i^2 = 44.50$

$$\frac{m(m^2-1)}{6[\sum d_i^2 + \frac{m(m^2-1)}{12} + \dots]}$$

$$\rho_{xy} = 1 - \frac{\dots}{\dots}$$

$$\frac{n(n^2-1)}{12}$$

where $d=R_1-R_2$ and 'm' is the number of times, an items is repeated.

Here $n=10$ and an item 86 is repeated twice i.e. $m=2$.

$$\frac{2(2^2-1)}{6[44.5 + \frac{2(2^2-1)}{12} + \dots]}$$

$$\rho_{xy} = 1 - \frac{12}{10 \times 99}$$

$$\frac{6(44.5+0.5)}{990} = \frac{6 \times 45}{990}$$

$$= 1 - \frac{6 \times 45}{990} = 1 - \frac{270}{990}$$

$$= \frac{990 - 270}{990} = \frac{720}{990} = 0.727.$$

There is high positive Correlation between x and y.

PARTIAL CORRELATION

It measures the relationship between any two variables when the other variables connected with those variables are kept constant.

It is a study of more than two variables one is D.V and others are I.V.

Examples:

- (i) The yield of crop may depend upon the rainfall, fertilizer, the average temperature, the period between sowing and harvesting.
- (ii) The value of house may depend upon cost, the locality, the tax rates etc.
- (iii) The IQ of children may depend upon their performances in Mathematics, English, the age etc.,

Problems

Q1. If $r_{12}=0.86$, $r_{13}=0.65$ and $r_{23}=0.72$ find the partial correlation coefficient.

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$\begin{aligned}
 & \frac{[0.86-(0.65 \times 0.72)]}{\sqrt{(1-0.65^2)(1-0.72^2)}} \\
 & = 0.7433.
 \end{aligned}$$

Q2. If $r_{12}=0.8$, $r_{13}= -0.4$ and $r_{23}=-0.56$, find partial correlation coefficient.

Solution:

Given

$$r_{12} = 0.8,$$

$$r_{13} = -0.4$$

$$r_{23} = -0.56,$$

$$\begin{aligned}
 r_{12.3} &= \frac{(r_{12}-r_{13}r_{23})}{\text{SQRT}[(1-r_{13}^2)(1-r_{23}^2)]}
 \end{aligned}$$

$$\begin{aligned}
 & \frac{[0.8-(0.4 \times 0.56)]}{\sqrt{[1-(-0.4)^2][1-(-0.56)^2]}} \\
 &= 0.759.
 \end{aligned}$$

$$\mathbf{r}_{13.2} = \frac{(\mathbf{r}_{13} - \mathbf{r}_{23} \mathbf{r}_{12})}{\sqrt{(1 - \mathbf{r}_{23}^2)(1 - \mathbf{r}_{12}^2)}}$$

$$\begin{aligned}
 & \frac{0.40 - (0.56 \times 0.8)}{\sqrt{[1 - (-0.56)^2][1 - (-0.8)^2]}} \\
 &= 0.097.
 \end{aligned}$$

$$\mathbf{r}_{23.1} = \frac{(\mathbf{r}_{23} - \mathbf{r}_{12} \mathbf{r}_{13})}{\sqrt{(1 - \mathbf{r}_{12}^2)(1 - \mathbf{r}_{13}^2)}}$$

$$\begin{aligned}
 & \frac{0.56 - [0.8 \times (-0.4)]}{\sqrt{[1 - (-0.8)^2][1 - (-0.4)^2]}} \\
 &=
 \end{aligned}$$

$$=0.436.$$

REGRESSION

Regression is the measure of the average relationship between two or more variables in terms of original units of data.

Example:

If the sales and advertisement are correlated we can find out expected amount of sales for a given advertising expenditure or the amount needed for attaining the given amount of sales.

Lines of regression

We shall have two regression lines as the regression line of X on Y and the regression line of Y on X.

The regression line of Y on X gives the most probable value of Y for given values of X and the regression line of X on Y gives the most probable values of X for given values of Y.

Formula:

Regression Equations:

(i) Equations of line of regression of Y on X

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

$$\sum(x - \bar{x})(y - \bar{y})$$

where $b_{yx} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$

$$\sum(x - \bar{x})^2$$

(ii) Equations of line of regression of X on Y.

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\sum(x - \bar{x})(\bar{y} - \bar{y})$$

where $b_{xy} = \frac{\sum (y - \bar{y})(x - \bar{x})}{\sum (y - \bar{y})^2}$

$$\sum (y - \bar{y})^2$$

Q4. From the following data, find

- (i) The two regression equations
- (ii) The coefficient of correlation between the marks in Economics and Statistics
- (iii) The most likely marks in statistics when marks in Economics are 30

Marks in Economics(x)	25	28	35	32	31	36	29	38	34	32
Marks in Statistics(y)	43	46	49	41	36	32	31	30	33	39

Solution:

x	y	$x - \bar{x}$ = x-32	$y - \bar{y}$ = y-38	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
						-93
320	380	0	0	140	398	-93

Here $\bar{x} = \frac{\sum x}{n}$

$\bar{y} = \frac{\sum y}{n}$

= 320 / 10

= 380 / 10

= 32

= 38

Equations of line of regression of Y on X

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

$$\sum(x - \bar{x})(y - \bar{y})$$

where $b_{yx} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$

$$\sum(x - \bar{x})^2$$

$$= -93 / 140$$

$$= -0.6643$$

Therefore $y - 38 = -0.6643(x - 32)$

$$y = -0.6643x + 38 + 0.6643 * 32$$

$$y = -0.6642x + 59.257$$

Equations of line of regression of X on Y.

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\sum(x - \bar{x})(y - \bar{y})$$

where $b_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(y - \bar{y})^2}$

$$\sum(y - \bar{y})^2$$

$$= -93 / 398$$

$$= -0.2337$$

Therefore $x - 32 = -0.2337(y - 38)$

$$x = -0.2337y + 40.88$$

coefficient of correlation

$$r^2 = b_{yx} * b_{xy}$$

$$= -0.6643 * (-0.2337)$$

$$= 0.1552$$

$$r = \sqrt{0.1552}$$

$$= 0.394$$

Now we have to find the most likely marks in statistics (y) when marks in economics (x) are 30.
We use the line of regression of y on x.

i.e. $y = -0.6643x + 59.2575$

put $x = 30$, we get $y = 39.32$

$$y = 39(\text{appr.})$$

The most likely marks in statistics (y) when marks in economics (x) are 30 calculated as 39.

CURVE FITTING

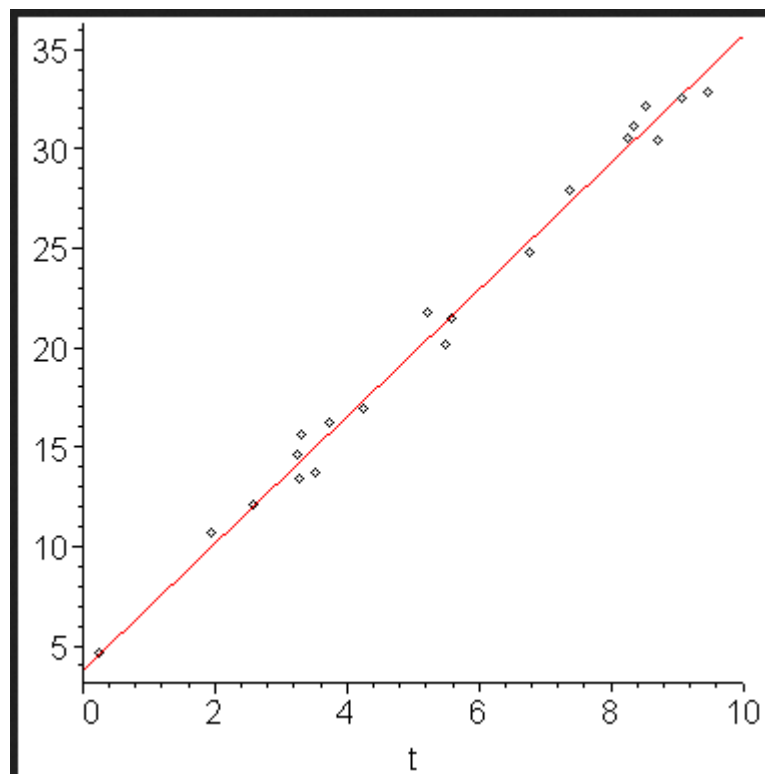
Curve fitting is the process of constructing a curve or a mathematical function that has the best fit to a set of points possibly subject to constraints.

For example, we measure the rainfall and yield of fields in Tamilnadu and represent those values by x_i and $y_i, i=1,2,3,\dots,n$.

We would like to know there is any relation between x and y . The empirical relation is written in the form of an equation $y = f(x)$.

Most of the time we may not able to get an exact relation but may get only approximate curve. If $(x_i, y_i), i=1,2,3,\dots,n$ are the n paired data which are plotted on the graph sheet, it is possible to draw a number of smooth curves passing through the points. The method of finding such approximating curve is called curve fitting.

$y=a+bx$ (To fit a straight line)



METHODS OF CURVE FITTING

1. The graphical methods
2. The Methods of Group Averages
3. The Method of Least Squares.

The first one is a rough method and in the second method evaluation of constants may vary. So we adopt another method called the method of least squares which gives a unique set of values to the constants in the equation of fitting curves.

METHOD OF LEAST SQUARES

The least squares method is a statistical procedure to find best fit for a set of data points by minimizing the sum of the residuals of points from the plotted curve.

TYPES OF CURVE

1. Fitting of a straight line : $y = a + bx$

$y \rightarrow$ Dependent Variable

$x \rightarrow$ Independent Variable

$a, b \rightarrow$ Constants

The normal equations are

$$\sum y = n a + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

2. Fitting of a parabolic curve: $y = a + bx + cx^2$

The normal equations are

$$\sum y = na + b \sum x + c \sum x^2$$

$$\sum xy = a \sum x + b \sum x^2 + c \sum x^3$$

$$\sum x^2 y = a \sum x^2 + b \sum x^3 + c \sum x^4$$

1. FITTING OF A STRAIGHT LINE

Let $y = a + bx$(1) be a straight line to be fitted to the given data.

Let (x_i, y_i) , $i = 1, 2, \dots, n$ be the n sets of observations, which fit the straight line (1) We have to select a and b which will best fit the straight line to the given data.

Working procedure:

- To fit the straight line $y = a + bx$
- Substitute the observed set of n values in this equation.
- Form the normal equations for each constant i.e.

The normal equations are

$$\sum y = n a + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

which are got by taking \sum on both sides of $y = a + bx$ and also taking \sum on both sides after multiplying by x both sides of equation (1).

Remark: Summing of constants n times will give n times of constant.

- Solve these normal equations as simultaneous equations of a and b .
Substitute the values of a and b in $y = a + bx$, which is required line of best fit.

Q1. Fit a straight line of the form $y = a + bx$ by using the methods

of least squares.

x	:	3	7	9	10
y	:	168	120	72	73

Solution:

Let the straight line be $y = a + bx$ (1)

The normal equations are

$$\sum y = n a + b \sum x \dots\dots\dots(2) \text{ and}$$

$$\sum xy = a \sum x + b \sum x^2 \dots\dots\dots(3)$$

x	y	xy	x^2
3	168	504	9
7	120	840	49
9	72	648	81
10	73	730	100
$\sum x = 29$	$\sum y = 433$	$\sum xy = 2722$	$\sum x^2 = 239$

Therefore (2) $\rightarrow 433 = 4a + 29b$ (here $n = 4$).....(4)

(3) $\rightarrow 2722 = 29a + 239b$ (5)

Multiply equation (4) by 29 and equation (5) by 4

$$116a + 841b = 12557$$

$$116a + 956b = 10888$$

Solving above two equations by changing sign we get,

$$-115b = 1669$$

$$b = -14.5$$

Substituting b value in equation (4) we get

$$4a + 29(-14.5) = 433$$

$$4a - 420.5 = 433$$

$$4a = 433 + 420.5$$

$$4a = 853.5$$

$$a = 213.375$$

Substituting a and b values in equation (1)

we get $y = 213.375 - 14.5x$ which is the required curve.

Q2. Fit a straight line $y = a + bx$, to the following data .

Year:	1991	1992	1993	1994	1995	1996	1997
Sales:	125	128	133	135	140	141	143
(in lakhs)							

Solution:

Let the straight line be $y = a + bx$(1)

The normal equations are

$$\sum y = na + b\sum x \dots\dots\dots(2) \text{ and}$$

$$\sum xy = a\sum x + b\sum x^2 \dots\dots\dots(3)$$

year	y(sales)	x=year – origin (1994)	x²	xy
1991	125	-3	9	-375
1992	128	-2	4	-256
1993	133	-1	1	-133
1994	135	0	0	0
1995	140	1	1	140

1996	141	2	4	282
1997	143	3	9	429
	$\sum y = 945$	$\sum x = 0$	$\sum x^2 = 28$	$\sum xy = 87$

i.e., $\sum x = 0$, $\sum y = 945$, $\sum x^2 = 28$, $\sum xy = 87$, $n = 7$

Substituting the above values in the normal equations we get,

$$945 = 7a + b(0) \dots \dots \dots (4)$$

$$87 = a(0) + b(28) \dots \dots \dots (5)$$

From (4) $a = 945 / 7 = 135$

From (5) $b = 87 / 28 = 3.11$

Therefore the straight line trend equation is $y = 135 + 3.11x$

Q3. Fit a straight line $y = a + bx$ to the following data.

Year : 1971 1972 1973 1974 1975 1976

Profit : 83 92 71 90 169 191

Solution:

Let the straight line be $y=a+bx$ (1)

The normal equations are

$$\sum y = n a + b \sum x \dots\dots\dots(2) \text{ and}$$

$$\sum xy = a \sum x + b \sum x^2 \dots\dots\dots (3)$$

Since $n=6$ (even), we take the origin to be 1973.5

Year	y(sales)	x=year – 1973.5	x^2	xy
1971	83	-2.5	6.25	-207.5
1972	92	-1.5	2.25	-138.0
1973	71	-0.5	0.25	-35.5
1974	90	0.5	0.25	45.0
1975	169	1.5	2.25	253.5
1976	191	2.5	6.25	477.5
Total	$\sum y = 696$	$\sum x = 0$	$\sum x^2 = 17.5$	$\sum xy = 395$

The normal equations are

$$696=6a+b(0)$$

$$395 =a(0)+b(17.5)$$

$$\text{Therefore, } a=696/6=116. \text{ } b=395/17.5=22.57$$

The straight line trend is given be $y=116+22.57x$.

FITTING OF A PARABOLA $y = a + bx + cx^2$

Let (x_i, y_i) , $i = 1, 2, \dots, n$ be set of observations of two variables x and y .

Let $y = a + bx + cx^2$ be the equation which fits best the given data.

The normal equations are

$$na + b\sum x + c\sum x^2 = \sum y \dots \dots \dots (1)$$

$$a\sum x + b\sum x^2 + c\sum x^3 = \sum xy \dots \dots \dots (2)$$

$$a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum x^2y \dots \dots \dots (3)$$

- (i) In $y = a + bx + cx^2$, take \sum on both sides
- (ii) Multiply by x both sides and then take \sum on both sides.
- (iii) Multiply both sides by x^2 and then take \sum on both sides.

Working Procedure:

- **Form Normal Equations**

$$na + b\sum x + c\sum x^2 = \sum y$$

$$a\sum x + b\sum x^2 + c\sum x^3 = \sum xy$$

$$a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum x^2y$$

- Solve these as simultaneous equations for a, b, c .
- Substitute the values of a, b, c in $y = a + bx + cx^2$ the required parabola of the best fit.

Q1. Fit a second degree parabola $y = a + bx + cx^2$ to the following data:

x:	1	2	3	4	5	6	7	8	9
y:	2	6	7	8	10	11	11	10	9

Solution:

Let the parabola be $y = a + bx + cx^2 \dots \dots (1)$

Whose normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \dots \dots \dots (2)$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \dots \dots \dots (3)$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4 \dots \dots \dots (4)$$

x	y	x ²	x ³	x ⁴	xy	x ² y
1	2	1	1	1	2	2
2	6	4	8	16	12	24
3	7	9	27	81	21	63
4	8	16	64	256	32	128
5	10	25	125	625	50	250
6	11	36	216	1296	66	396
7	11	49	343	2401	77	539
8	10	64	512	4096	80	640
9	9	81	729	6561	81	729
$\sum x = 45$	$\sum y = 74$	$\sum x^2 = 285$	$\sum x^3 = 2025$	$\sum x^4 = 15333$	$\sum xy = 421$	$\sum x^2y = 2771$

Therefore (2) $\rightarrow 74 = 9a + 45b + 285c \dots \dots \dots (5)$ (since $n = 9$)

$$(3) \rightarrow 421 = 45a + 285b + 2025c \dots \dots \dots (6)$$

$$(4) \rightarrow 2771 = 285a + 2025b + 15333c \dots \dots (7)$$

Solving Eqn (5) and (6) by multiplying Eqn.(5) by 5 and solving Eqn (6) and (7) by multiplying Eqn.(6) by 285 and Eqn (7) by 45.

$$60b + 600c = 51 \dots \dots \dots (8)$$

$$220b + 2508c = 104.67 \dots \dots \dots (9)$$

Solving Eqn(8) and (9) by multiplying Eqn (8) by 220 and Eqn (9) by 60

$$\rightarrow 18480c = -4939.8$$

$$\rightarrow c = -0.2673$$

Substituting c value in Eqn.(8)

$$60b = 51 - 600(-0.2673)$$

$$60b = 211.38$$

$$\rightarrow b = 3.523$$

Now by substituting b and c values in Eqn.(5)

$$a = -0.9283$$

Therefore, $y = -0.9283 + 3.523x - 0.2673x^2$ is the required parabola.

Q2. Fit a second degree polynomial equation to the following data.

Year(x):	1976	1977	1978	1979	1980	1981	1982	1983	1984
Sales(y):	50	65	70	85	82	75	65	90	95
(in lakhs)									

Solution:

The second degree polynomial equation is

$$y = a + bx + cx^2 \dots\dots\dots (1)$$

The normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \dots\dots\dots (2)$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \dots\dots\dots (3)$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4 \dots\dots\dots (4)$$

Here n=9.

year	y	x=year-1980	x ²	x ³	x ⁴	xy	x ² y
1976	50	-4	16	-	256	-200	800
1977	65	-3	9	64	81	-195	585
1978	70	-2	4	-	16	-140	280
1979	85	-1	1	27	1	-85	85
1980	82	0	0	-8	0	0	0
1981	75	1	1	-1	1	75	75
1982	65	2	4	0	16	130	260
1983	90	3	9	1	81	270	810
1984	95	4	16	8	256	380	1520
				27			
				64			
Total	$\sum y = 677$	$\sum x = 0$	$\sum x^2 = 60$	$\sum x^3 = 0$	$\sum x^4 = 70$	$\sum xy = 235$	$\sum x^2y = 4415$

Substituting above values we get

$$677 = 9a + 60c \dots\dots\dots (5)$$

$$235 = 60b \dots\dots\dots (6)$$

$$4415 = 60a + 708c \dots\dots (7)$$

Solving we get

$$a = 77.3509, b = 3.9167, c = -0.3193$$

Hence, the second degree polynomial trend equation is

$$y = 77.3509 + 3.9167x - 0.3193x^2.$$

Q3. The price of a commodity during 1993-98 are given below. Fit a parabola $y = a + bx + cx^2$ to these data. Calculate the trend values. Estimate the price of commodity for the year 1999.

Year:	1993	1994	1995	1996	1997	1998
Price:	100	107	128	140	181	192.

Solution:

The required parabola is $y=a+bx+cx^2$ (1)

The normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \dots\dots\dots(2)$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \dots\dots\dots(3)$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4 \dots\dots\dots(4)$$

Year	Price y	x=year- -1995.5	x^2	x^3	x^4	xy	x^2y
1993	100	-2.5	6.25	-15.625	39.062	-250	625
1994	107	-1.5	2.25	-3.375	5.0625	-160.5	240.75
1995	128	-0.5	0.25	-0.125	0.0625	-64	32
1996	140	0.5	0.25	0.125	0.0625	70	35
1997	181	1.5	2.25	3.375	5.0625	271.5	407.2
1998	192	2.5	6.25	15.625	39.0625	480	1200
<hr/>							
$\sum y = 848$		$\sum x = 0$	$\sum x^2 = 17.5$	$\sum x^3 = 0$	$\sum x^4 = 88.37$	$\sum xy = 347$	$\sum x^2y = 2540$
<hr/>							

The normal equations are

$$848 = 6a + 17.5c \dots\dots\dots(5)$$

$$347 = 17.5b \dots\dots\dots(6)$$

$$2540 = 17.5a + 88.37c \dots\dots\dots(7)$$

Solving,

$$a = 136.12, b = 19.83, c = 1.786$$

Hence required parabola is $y = 136.12 + 19.83x + 1.786x^2$.

The trend values are calculated in the table

For the year 1999, $x = 1999 - 1995.5 = 3.5$

Therefore, Price in 1999 = $136.12 + (19.83 \times 3.5) + (1.786 \times 3.5 \times 3.5)$

Rs.227.4035

FITTING OF THE CURVE OF THE FORM $y = a + bx_1 + cx_2$

The normal equations are

$$\begin{aligned}\sum y &= na + b\sum x_1 + c\sum x_2 \\ \sum x_1 y &= a\sum x_1 + b\sum x_1^2 + c\sum x_1 x_2 \\ \sum x_2 y &= a\sum x_2 + b\sum x_1 x_2 + c\sum x_2^2\end{aligned}$$

Q1. Fit the curve of the form $y = a + bx_1 + cx_2$ to the following data:

y	7	12	17	20
x ₁	4	7	9	12
x ₂	1	2	5	8

Solution:

Let $y = a + bx_1 + cx_2$ be the best fit of the given data.

The normal equations are

$$\begin{aligned}\sum y &= na + b\sum x_1 + c\sum x_2 \\ \sum x_1 y &= a\sum x_1 + b\sum x_1^2 + c\sum x_1 x_2 \\ \sum x_2 y &= a\sum x_2 + b\sum x_1 x_2 + c\sum x_2^2\end{aligned}$$

x ₁	x ₂	y	x ₁ ²	x ₂ ²	x ₁ x ₂	x ₁ y	x ₂ y
4	1	7	16	1	4	28	7
7	2	12	49	4	14	84	24
9	5	17	81	25	45	153	85
12	8	20	144	64	96	240	160
32	16	56	290	94	159	505	276
$\sum x_1$	$\sum x_2$	$\sum y$	$\sum x_1^2$	$\sum x_2^2$	$\sum x_1 x_2$	$\sum x_1 y$	$\sum x_2 y$

$$(2) \rightarrow 4a + 32b + 16c = 56$$

$$(3) \rightarrow 32a + 290b + 159c = 505$$

$$(4) \rightarrow 16a + 159b + 94c = 276$$

Solving we get $a = 0.6444$, $b = 1.661$, $c = 0.0169$.

Therefore the required equation is

$$y = a + bx_1 + cx_2: y = 0.6444 + 1.661x_1 + 0.0169x_2.$$

UNIT II QUESTION BANK

PART-A

1. State the properties of correlation coefficient.
2. Outline the formulae for Regression Equations.
3. Find the Straight line and hence find y when X = 31
X: 18 19 20 21 22 23 24 25 26 27
Y: 17 17 18 18 11 19 19 20 21 22
4. Lists the types of correlation.
5. Define Scatter Plot.
6. The following are the ranks obtained by 10 students in statistics and mathematics. To what extent is knowledge of students in statistics related to knowledge in mathematics?

Rank of Stats	: 1	2	3	4	5	6	7	8	9	10
Rank of Maths	:2	4	1	5	3	9	7	10	6	8

7. Define curve fitting.
8. State principle of Least square method.
9. Lists the formula for fitting straight line $y=a+bx$
10. Lists the formula for fitting second degree parabola $y=ax^2+bx+c$.

PART-B

1. Determine the rank correlation coefficient of correlation of the following data:

X: 80 78 75 75 68 67 60 59

Y: 12 13 14 14 14 16 15 17

2. From the data given below Compose (i) two regression lines (ii) coefficient of correlation between marks in Physics and marks in Chemistry (iii) most likely marks in Chemistry when marks in Physics is 78 (iv) most likely marks in Physics when marks in Chemistry is 92

Marks in Physics(X):72 85 91 85 91 89 84 87 75 77

Marks in Chem (Y):76 92 93 91 93 95 88 91 80 81

3. A computer while calculating the correlation coefficient between two variates x and y from 25 pairs of observations obtained the following constants $N = 25$, $\sum XY = 516$, $\sum X = 125$, $\sum Y = 100$, $\sum X^2 = 650$, $\sum Y^2 = 480$. It was however detected later on at the time of checking that he

copied down pairs (6,4) and (8,6) for (8,12) and (6,8). Obtain the correct value of the correlation coefficient.

4. In an aptitude test 2 judges rank the 10 competitors in the following order

Judge I : 6 4 3 2 1 7 9 8 10 5

JudgeII : 4 1 6 7 5 8 10 9 3 2

Is there any concordance between the two judges?

5. Obtain the coefficient of correlation between x and y from the following

X 10 12 13 16 17 20 25

y 19 22 26 27 29 33 37

6. Fit a straight line of the form $y = ax + b$ by using least squares method:

X: 0 5 10 15 20 25

Y: 12 15 17 22 24 30

7. Find the regression line of y on x

X	1	4	2	3	5
Y	3	11	2	15	4

8. The following data give the height in inches(x) and the weight in pounds (y) of random sample of 10 students from a large group of age 17 years:

x: 61 68 68 64 65 70 63 62 64 67

y: 112 123 130 115 110 125 100 113 116 126

Find two regression equations and hence find r.

9. Three judges rank the 10 entries in a beauty contest as follows :

X : 8 7 5 4 9 10 6 2 1 3

Y : 7 8 9 3 10 6 5 4 2 1

Z : 10 9 8 7 6 5 3 4 1 2

Using rank correlation coefficient method, discuss which pair of judges has the nearest approach to common taste and beauty.

10. Fit a parabola of the form $y = ax^2 + bx + c$ for the following data:

x:	1	2	3	4	5
y:	3	4	7	12	21

UNIT III BASIC STATISTICS – SMTA1207

PROBABILITY THEORY & RANDOM VARIABLES

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, $PV = \text{constant}$.

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 of 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,

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

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: $0 \leq P(E) \leq 1$

Axiom 2: $P(S) = 1$

Axiom 3: For any sequence of mutually exclusive events, E_1, E_2, \dots (that is, events for which $E_i \cap E_j = \Phi$, 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, \dots, A_n are mutually exclusive, then

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

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	$\omega \in \bar{A} \cap \bar{B}$
4. Event A occurs and B does not occur	$\omega \in A \cap \bar{B}$
5. Exactly one of the events A or B occurs	$\omega \in A \Delta B$
6. If event A occurs, so does B	$A \subset B$
7. Events A and B are mutually exclusive	$A \cap B = \Phi$
8. Complementary event of A	\bar{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 $\frac{1}{2}$.

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

∴ The required probability is $\frac{1}{4}$.

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

∴ The required probability is $= \frac{3}{6} = \frac{1}{2}$.

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 ${}^{16}C_2$ 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 ${}^6C_1 \times {}^7C_1 = 6 \times 7 = 42$.

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

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.

$$\begin{aligned} \text{(i)} \quad P(\text{both are good}) &= \frac{\text{No. of ways drawing 2 good articles}}{\text{Total no. of ways of drawing 2 articles}} \\ &= \frac{{}^{10}C_2}{{}^{16}C_2} = \frac{3}{8} \end{aligned}$$

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

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

$$\text{(iii)} \quad P(\text{at least 1 is good}) = P(\text{exactly 1 is good or both are good})$$

$$= P(\text{exactly 1 is good and 1 is bad or both are good})$$

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

(iv) $P(\text{atmost 1 is good}) = P(\text{none is good or 1 is good and 1 is bad})$

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

(v) $P(\text{exactly 1 is good}) = P(1 \text{ is good and 1 is bad})$

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

(vi) $P(\text{neither has major defects}) = P(\text{both are non-major defective articles})$

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

(vii) $P(\text{neither is good}) = P(\text{both are defective})$

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

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

$$= 6C_2 \times 8C_2 = 420.$$

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

$$= 14C_4 = 1001.$$

$P(\text{the product is positive})$

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

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

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 $P(Pi \cap W) = 0.67 \cdot 0.30 = 0.20$. Utilizing Formula 2.9,

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

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* genotype 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 $OoYyPp \times OoYypp$, 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:

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

Phenotypic array:

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

The probability 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

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

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

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

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 ${}^{52}C_2$ ways, all being equally likely.

\therefore Exhaustive number of cases is ${}^{52}C_2$.

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

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

(b) Exhaustive number of cases = ${}^{52}C_3$

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 favourable cases = ${}^4C_1 \times {}^4C_1 \times {}^4C_1$.

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

(c) Exhaustive number of cases ${}^{52}C_4$

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

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

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

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 ${}^{52}C_{13}$

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

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

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 ${}^{10}C_4$ ways. Hence exhaustive number of cases is ${}^{10}C_4 = 210$

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

$$\therefore \text{Required probability} = \frac{24}{210}$$

- (ii) $P(\text{Committee has at least one purchase officer}) = 1 - P(\text{Committee has no purchase Officer})$

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 ${}^6C_4 = 15$ ways. Hence,

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

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

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

$$1 \times {}^9C_3 = 84 \text{ ways.}$$

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 9C_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,

$$\begin{aligned} P(E) &= P(i) + P(ii) + P(iii) \\ &= \frac{{}^6C_1 \times {}^4C_1 \times {}^5C_2}{{}^{15}C_4} + \frac{{}^6C_2 \times {}^4C_1 \times {}^5C_1}{{}^{15}C_4} + \frac{{}^6C_1 \times {}^4C_2 \times {}^5C_1}{{}^{15}C_4} \\ &= 0.5275 \end{aligned}$$

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

$$\begin{aligned} P(A) &= 1/2 & P(B) &= 3/4 & P(C) &= 1/4 \\ P(\bar{A}) &= 1 - 1/2 = 1/2 & P(\bar{B}) &= 1 - 3/4 = 1/4 & P(\bar{C}) &= 1 - 1/4 = 3/4 \end{aligned}$$

$$P(\text{Problem solved}) = P(\text{At least one of them solves the problem})$$

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

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

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

$$\begin{aligned}
&= 1 - P(\bar{A}) P(\bar{B}) P(\bar{C}) \\
&= 1 - \frac{1}{2} \times \frac{1}{4} \times \frac{3}{4} \\
&= \frac{29}{32}
\end{aligned}$$

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,

$$\text{Required probability} = P(i) + P(ii) + P(iii)$$

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

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

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

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

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

Conditional Probability and Baye's Theorem

Conditional Probability and Multiplication Law

For two events A and B

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B/A), P(A) > 0 \\ &= P(B) \cdot P(A/B), P(B) > 0 \end{aligned}$$

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, \dots, 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.

$$\begin{aligned} P(A \cap B) &= P(\text{both tubes drawn are good}) \\ &= \frac{{}^6C_2}{{}^{10}C_2} = \frac{1}{3} \end{aligned}$$

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.

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

$P(D/A) = P(\text{an item is defective, given that A has produced it})$

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

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

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.

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

By theorem of total probability,

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

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 placed 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.

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

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

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

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

$$= P(\text{drawing a white ball} / \text{urn II contains 5 white and 5 black balls})$$

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

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

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 probability 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,

$$P(E_1) = \frac{4}{9}$$

$$P(E_2) = \frac{2}{9}$$

$$P(E_3) = \frac{3}{9}$$

$$P(A/E_1) = 0.3$$

$$P(A/E_2) = 0.5$$

$$P(A/E_3) = 0.8$$

$$P(A) = P[(A \cap E_1) \cup (A \cap E_2) \cup (A \cap E_3)]$$

$$= P[(A \cap E_1) + (A \cap E_2) + (A \cap E_3)]$$

$$= P(E_1) P(A/E_1) + P(E_2) P(A/E_2) + P(E_3) P(A/E_3)$$

$$= \frac{4}{9} \times \frac{3}{10} + \frac{2}{9} \times \frac{5}{10} + \frac{3}{9} \times \frac{8}{10} = \frac{23}{45}$$

3.3.4. Baye's theorem

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

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

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.

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

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

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,

$$\begin{aligned}
 P(B_4/A) &= \frac{P(B_4) \times P(A/B_4)}{\sum_{i=1}^4 P(B_i) \times P(A/B_i)} \\
 &= \frac{\frac{1}{4} \times 1}{\frac{1}{4} \times \left(\frac{1}{10} + \frac{3}{10} + \frac{3}{5} + 1 \right)} = \frac{1}{2}
 \end{aligned}$$

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 been 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

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

By Baye’s theorem

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

$$\begin{aligned}
 &= \frac{\frac{1}{4} \times 1}{\frac{1}{4} \times 1 + \frac{3}{4} \times \frac{1}{16}} = \frac{16}{19}
 \end{aligned}$$

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

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

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

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

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

$$\begin{aligned} \text{Hence } P(E_2/A) &= \frac{P(E_2)P(A/E_2)}{\sum_{i=1}^3 P(E_i)P(A/E_i)} \\ &= \frac{\frac{1}{3} \times \frac{1}{3}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{11}} = \frac{55}{118} \end{aligned}$$

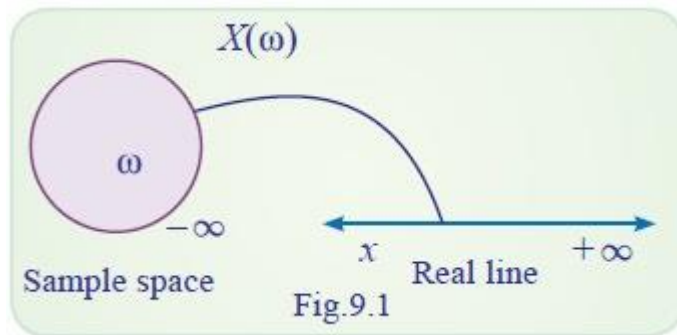
$$\text{Similarly, } P(E_3/A) = \frac{\frac{1}{3} \times \frac{2}{11}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{11}} = \frac{30}{118}$$

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

Definition of random variable

Definition

Let S be the sample space of a random experiment. A rule that assigns a single real number to each outcome (sample point) of the random experiment is called random variable.



In other words, a random variable is a real valued function defined on a sample space S that is with each outcome ω of a random experiment there corresponds a unique real value x known as a value of the random variable X . That is $X(\omega) = x$.

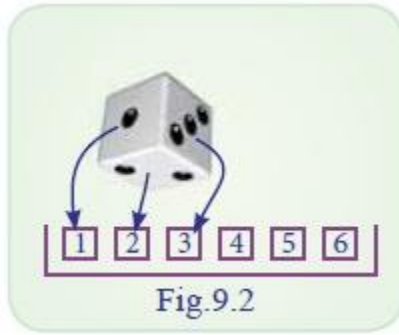
Generally random variables are denoted by upper case alphabets like $X, Y, Z \dots$ and their values or realizations are denoted by the corresponding lower case letters. For example, if X is a random variable, the realizations are $x_1, x_2 \dots$

Example 1

Consider the random experiment of rolling a die.

The sample space of the experiment is $S = \{1, 2, 3, 4, 5, 6\}$

Let X denotes the face of the die appears on top.



The assigning rule is

$$X(1) = 1, X(2) = 2, X(3) = 3, X(4)=4, X(5)=5 \text{ and } X(6)=6$$

Hence the values taken by the random variable X are 1,2,3,4,5,6. These values are also called the realization of the random variable X .

Example 2

Random experiment : Two coins are tossed simultaneously.

Sample space : $S=\{HH, HT, TH, TT\}$

Assigning rule : Let X be a random variable defined as the number of heads comes up.

Sample Point ω	HH	HT	TH	TT
$X(\omega)$	2	1	1	0

Here, the random variable X takes the values 0, 1, 2 .

Example 3

Experiment : Two dice are rolled simultaneously.

Sample space : $\{(1, 1),(1, 2),(1, 3),\dots(6, 6)\}$

Assigning rule : Let X denote the sum of the numbers on the faces of dice

then $X_{ij} = i + j$, Here, i denotes face number on the first die and j denotes the face number on the second die.

Then X is a random variable which takes the values 2, 3, 4 12.

That is the range of X is $\{2, 3, 4, \dots, 12\}$

Discrete and Continuous random variables

Random variables are generally classified into two types, based on the values they take such as Discrete random variable and Continuous random variable.

Discrete random variable

A random variable is said to be discrete if it takes only a finite or countable infinite number of values.

Example 4

Consider the experiment of tossing a coin

If X (Head) = 1, X (Tail) = 0

Then X takes the values either 0 or 1

This is a discrete random variable.

Example 5

Consider the experiment of tossing a coin till head appears.

Let random variable X denote the number of trials needed to get a head. The values taken by it will be 1, 2, 3, ..

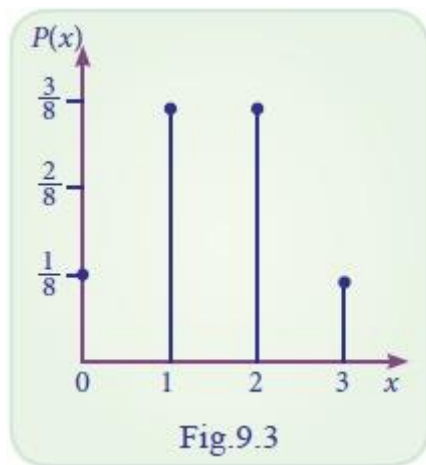
It is discrete random variable taking countable infinite values.

Continuous random variable:

A random variable X is said to be continuous, if it takes values in an interval or union of disjoint intervals. (A rigorous definition is beyond the scope of the book).

Probability mass function and probability density function

A probability function is associated with each value of the random variable. This function is used to compute probabilities for events associated with the random variables. The probability function defined for a discrete random variable is called probability mass function. The probability function associated with continuous random variable is called probability density function.



Probability Mass Function

If, X is a discrete random variable taking values x_1, x_2, \dots, x_n with respective probabilities $p(x_1), p(x_2), \dots, p(x_n)$ such that

$$(i) p(x_i) \geq 0, \quad \forall i \text{ (non-negative)}$$

$$\text{and (ii) } \sum_{i=1}^n p(x_i) = 1$$

then $p(x)$ is known as the probability mass function (p.m.f) of the discrete random variable X .

The pair $\{x_i, p(x_i); i = 1, 2, 3, \dots\}$ is known as probability distribution of X .

Example 8

A coin is tossed two times. If X is the number of heads, find the probability mass function of X .

Solution:

Since the coin is tossed two times, the sample space is $S=\{HH, HT, TH, TT\}$

If X denotes the numbers of heads, the possible values of X are 0,1,2 with the following

$$P(X = 0) = P(\text{getting no head}) = \frac{1}{4}$$

$$P(X = 1) = P(\text{getting one head}) = \frac{2}{4} = \frac{1}{2}$$

$$P(X = 2) = P(\text{getting two heads}) = \frac{1}{4}$$

The probability distribution of X is

X	0	1	2
$p(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Example 9

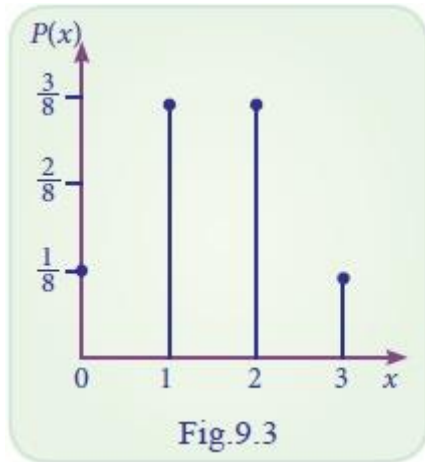
In example 9.3 the probability mass function of X is given in the following table

X	2	3	4	5	6	7	8	9	10	11	12
$P(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The above table may be called as the probability distribution function of X .

Probability mass function and probability density function

A probability function is associated with each value of the random variable. This function is used to compute probabilities for events associated with the random variables. The probability function defined for a discrete random variable is called probability mass function. The probability function associated with continuous random variable is called probability density function.



Probability Density Function

When the random variable is continuous in the co-domain it is spread over it. A function $f(x)$ is defined on real line and satisfying the following conditions :

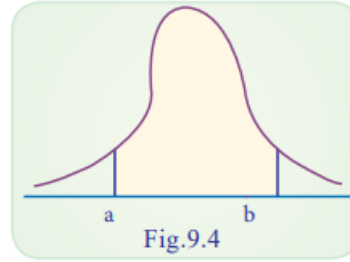
(i) $f(x) \geq 0, \forall x$

(ii) $\int_{-\infty}^{\infty} f(x) dx = 1$

is called the probability density function(p.d.f) of X.

Remark:

- (i) Every integrable function satisfying the above two conditions is a probability density function of some random variable X
- (ii) The probability that the value of X lies between two points ' a ' and ' b ' is $P(a < X < b) = \int_a^b f(x) dx$
- (iii) If X is discrete random variable then for any real x , $P(X = x)$ need not be zero. However in the case of continuous random variable $P(X = x) = 0$ holds always. $P(X = a) = \int_a^a f(x) dx = 0$
- (iv) If X is a continuous random variable then for any $a < b$
 $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$



Example 10

A continuous random variable X has probability density function given by

$$f(x) = \begin{cases} Ax^3, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}. \text{ Find A.}$$

Solution:

Since $f(x)$ is a p.d.f

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$\int_0^1 Ax^3 dx = 1$$

$$\Rightarrow \frac{A}{4} = 1$$

$$\Rightarrow A = 4$$

Example 11

Verify whether the following function is a probability density function

$$f(x) = \begin{cases} \frac{2x}{9}, & 0 < x < 3 \\ 0, & \text{elsewhere} \end{cases}$$

Solution:

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^3 \frac{2x}{9} dx$$

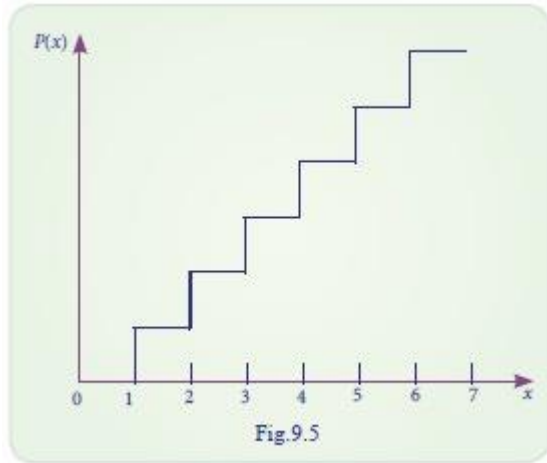
$$\Rightarrow \frac{2}{9} \left[\frac{x^2}{2} \right]_0^3 = 1$$

It is to be noted that (i) $f(x) \geq 0, \forall x$ (ii) $\int_{-\infty}^{\infty} f(x) dx = 1$

Hence, $f(x)$ is a p.d.f.

Distribution function and its properties

We get the probability of a given event at a particular point. If we want to have the probability upto the point we get the probability $P(X \leq x)$. This type of probability is known as probability mass function. We can also find how the probability is distributed within certain limits. $[P(X < x)$ or $P(X > x)$ or $P(a < x < b)]$.



Distribution Function for discrete random variable

Definition: Let X be a random variable, the cumulative distribution function (c.d.f) of a random variable X is defined as $F(x) = P(X \leq x)$, $\forall x$. It is called simply as distribution function.

Properties:

- (i) $0 \leq F(x) \leq 1, \forall x$, (non -negative)
- (ii) $F(x) \leq F(y), \forall x < y$, (non -decreasing)
- (iii) $\lim_{h \rightarrow 0} F(x+h) = F(x), \forall x$, ($F(x)$ is right continuous)
- (iv) $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$
- (v) $F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1$

Distribution Function for continuous random variable

- (i) $F(x) = \int_{-\infty}^x f(x) dx$
- (ii) $f(x) = F'(x)$
- (iii) $P(a < X < b) = P(a \leq X < b)$
 $= P(a < X \leq b) = P(a \leq X \leq b)$
 $= \int_a^b f(x) dx$

Properties

- (i) $F(x)$ is a non decreasing function of x
- (ii) $0 < F(x) < 1 \quad -\infty < x < \infty$
- (iii) $F(-\infty) = 0,$
- (iv) $F(\infty) = 1$
- (v) For any real constant a and b such that $a < b$, $p(a < X \leq b) = F(b) - F(a)$
- (vi) $f(x) = \frac{d}{dx}(F(x)) \quad \text{i.e., } f(x) = F'(x)$

Example 12

A random variable X has the following probability mass function

X	0	1	2	3	4	5	6
$P(X = x)$	a	$3a$	$5a$	$7a$	$9a$	$11a$	$13a$

- i. Find the value of ' a '
- ii. Find the c.d.f $F(x)$ of X
- iii. Evaluate : (a) $P(X \geq 4)$ (b) $P(X < 5)$ (c) $P(3 \leq X \leq 6)$
- iv. $P(X = 5)$ using $F(x)$

Solution:

(i) Since $P(X = x)$ is probability mass function $\sum P(X = x) = 1$

i.e., $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) = 1$

$$a + 3a + 5a + 7a + 9a + 11a + 13a = 1$$

$$49a = 1 \Rightarrow a = \frac{1}{49}$$

(ii) Hence the c.d.f is

X	0	1	2	3	4	5	6
$P(x)$	$\frac{1}{49}$	$\frac{3}{49}$	$\frac{5}{49}$	$\frac{7}{49}$	$\frac{9}{49}$	$\frac{11}{49}$	$\frac{13}{49}$
$F(x)$	$\frac{1}{49}$	$\frac{4}{49}$	$\frac{9}{49}$	$\frac{16}{49}$	$\frac{25}{49}$	$\frac{36}{49}$	$\frac{49}{49} = 1$

$$(iii) \quad (a) \quad P(X \geq 4)$$

$$= P(X = 4) + P(X = 5) + P(X = 6)$$

$$= 9a + 11a + 13a$$

$$= 33a$$

$$= 33 \times \frac{1}{49} = \frac{33}{49}$$

$$(b) \quad P(X < 5) = 1 - P(X \geq 5) = 1 - [P(X = 5) + P(X = 6)]$$

$$= 1 - [11a + 13a]$$

$$= 1 - 24a$$

$$= 1 - \frac{24}{29}$$

$$= \frac{25}{29}$$

$$(c) \quad P(3 \leq X \leq 6) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$$

$$= 7a + 9a + 11a + 13a$$

$$= 40$$

$$a = \frac{40}{49}$$

$$(iv) \quad P(X = 5) = F(5) - F(5 - 1)$$

$$= \frac{36}{49} - \frac{25}{49}$$

$$= \frac{11}{49}.$$

Example 13

Let X be a random variable with p.d.f

$$f(x) = \begin{cases} \frac{x}{2} ; & 0 < x < 2 \\ 0 ; & \text{otherwise} \end{cases}$$

- (i) Find the c.d.f of X ,
- (ii) Compute $P\left(\frac{1}{2} < X \leq 1\right)$,
- (iii) Compute $P(X=1.5)$

(i) The c.d.f of X : $F(x) = P(X \leq x)$

$$= \int_{-\infty}^x f(x) dx$$

$$= \int_0^x \frac{x}{2} dx = \frac{x^2}{4}$$

$$\text{Hence, } F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x^2}{4} & \text{if } 0 < x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

$$\begin{aligned} \text{(ii)} \quad P\left(\frac{1}{2} < X \leq 1\right) &= F(1) - F\left(\frac{1}{2}\right) \\ &= \frac{1}{4} - \frac{1}{16} = \frac{3}{16} \end{aligned}$$

This probability can be computed using p.d.f

$$\begin{aligned} \int_{\frac{1}{2}}^1 \frac{x}{2} dx &= \frac{1}{2} \left(\frac{x^2}{2} \right)_{\frac{1}{2}}^1 \\ &= \frac{3}{16} \end{aligned}$$

$$\text{(iii)} \quad P(X = 1.5) = 0$$



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

UNIT III QUESTION BANK

PART-A

1. A bag contains 5 white balls and 3 black balls .Two balls are drawn at random one after the other without replacement. Find the probability that both balls drawn are black.
2. Define Baye's theorem
3. State Conditional probability
4. Define Addition theorem of probability.
5. State Multiplication theorem of probability.
6. Find the probability of getting two tails in two tosses of a coin.
7. What is the chance that a leap year selected at random will contain 53 Sundays?
8. State Probability of mass function.
9. State probability density function
10. Write the properties of distribution function.

PART B

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

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?

3. 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?

4.

Values of X	0	1	2	3	4	5	6	7	8
Probability P(x)	a	3a	5a	7a	9a	11a	13a	15a	17a

- (i) Determine the value of a
 - (ii) Find $P(X < 3)$, $P(X \geq 3)$, $P(0 < X < 5)$
5. The diameter of an electric cable is assumed to be a continuous random variable with p.d.f $f(x) = 6x(1-x)$, $0 \leq x \leq 1$
- (i) Check that above is a p.d.f.
 - (ii) Determine a number “b” such that $P(X < b) = P(X > b)$
6. In a bolt factory machines A, B, C manufacture respectively 25%, 35% and 40% of the total output. Of their output 5%, 4% and 2% are defective bolts. A bolt is drawn at random from the product and is found to be defective. What are the probabilities that it was manufactured by machines A, B and C.
7. A box contains 6 red, 4 white and 5 black balls. A person 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.
8. 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 placed 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?
- 9.

Unit IV BASIC STATISTICS – SMTA1207

STANDARD DISTRIBUTIONS

Introduction

The probability of the various values of the random variables are obtained in accordance with the events and the nature of the experiment

In this chapter we are going to see some distributions called theoretical distributions. In these distributions, probabilities of the events are to be obtained using (formula) derived under certain conditions or assumptions. Of the many distributions available, the more common are Bernoulli, Binomial, Poisson, Normal, and Uniform distributions.

In practical situations one has to thoroughly understand the random environment and to describe it. It is followed by suggesting one of the above probability functions suitable to the situation and to obtain the requirement. The characteristics of the probability distributions such as Central Tendency, Dispersions, and Skewness are also to be studied.

Discrete distributions

Bernoulli's Distribution

It is discovered by a Swiss Mathematician James Bernoulli (1654-1705) for a trial which has only two outcomes viz. a success with probability p and a failure with probability $q = 1 - p$.

Definition

A random variable X is said to follow a Bernoulli distribution if its probability mass function is given by

$$P(X = x) = \begin{cases} p^x q^{1-x} & x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

Characteristics of Bernoulli Distribution

- i. Number of trials is one
- ii. $q = 1 - p$
- iii. Constants of the distributions
- iv. (i) mean = p (ii) variance = pq (iii) standard deviation = \sqrt{pq}

BINOMIAL DISTRIBUTION

Introduction

Binomial distribution was discovered by James Bernoulli (1654 – 1705) in the year 1700 and was first published in 1713 eight years after his death. The distribution of the Sum of n independent Bernoulli variables is known as a Binomial distribution.

That is the sum of outcome of n independent experiments of Bernoulli trials, in each of which the probability of success is constant p and the probability of failure is $q=1-p$ is called Binomial experiment.

Definition

A random variable X denoting the number of successes in an outcome of a Binomial experiment having n trials and p as the probability of success in each trial is called Binomial random variable. Its probability mass function is given by

$$P(X = x) = \begin{cases} nC_x p^x q^{n-x} & \text{for } x = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases} \quad \text{where } q=1-p$$

Conditions for Binomial Distribution

We get the Binomial Distribution under the following experimental conditions:

- i. The number of trials ' n ' is finite
- ii. The trials are independent of each other
- iii. The probability of success ' p ' is same for each trial
- iv. Each trial must result in a success or a failure.

Characteristics of Binomial Distribution

i. Binomial distribution is a discrete distribution i.e., X can take values $0, 1, 2, \dots, n$ where ' n ' is finite .

ii. Constants of the distributions are:

Mean = np ; Variance = npq ; Standard deviation = \sqrt{npq}

$$\text{Skewness} = \frac{q - p}{\sqrt{npq}} ; \quad \text{Kurtosis} = \frac{1 - 6pq}{npq}$$

iii. It may have one or two modes.

iv. If $X \sim B(n_1, p)$ and $Y \sim B(n_2, p)$ and that X and Y are independent then $X + Y \sim B(n_1 + n_2, p)$

v. If ' n ' independent trials are repeated N times the expected frequency of ' x ' successes is $N \times {}^nC_x p^x q^{n-x}$

vi. If $p = 0.5$, the distribution is symmetric.

Example 1

Comment on the following 'The mean of binomial distribution is 5 and its variance is 9'.

Solution:

Given mean $np = 5$ and variance $npq = 9$

$$\therefore \frac{\text{Value of variance}}{\text{Value of mean}} = \frac{npq}{np} = \frac{9}{5} \therefore q = \frac{9}{5} > 1 \quad \text{is not possible}$$

as $0 \leq q \leq 1$ and hence the given statement is wrong.

Example 2

Eight coins are tossed simultaneously. Find the probability of getting atleast six heads.

Solution:

$$\text{Here } n=8 \quad p = P(\text{head}) = \frac{1}{2} \quad q = 1 - \frac{1}{2} = \frac{1}{2}$$

Trials satisfy conditions of Binomial distribution

$$\begin{aligned} \text{Hence } P(X = x) &= {}^nC_x p^x q^{n-x} \quad x = 0, 1, 2, \dots, n \\ &= {}^8C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{8-x} \quad x = 0, 1, 2, \dots, 8 \\ &= {}^8C_x \left(\frac{1}{2}\right)^{x+8-x} \\ &= {}^8C_x \left(\frac{1}{2}\right)^8 \end{aligned}$$

$$\therefore P(X = x) = \frac{{}^8C_x}{256}$$

$P(\text{getting atleast six heads})$

$$\begin{aligned} &= P(x \geq 6) \\ &= P(x = 6) + P(x = 7) + P(x = 8) \\ &= \frac{{}^8C_6}{256} + \frac{{}^8C_7}{256} + \frac{{}^8C_8}{256} \\ &= \frac{28}{256} + \frac{8}{256} + \frac{1}{256} \\ &= \frac{37}{256} \end{aligned}$$

Example 3

Ten coins are tossed simultaneously. Find the probability of getting (i) atleast seven heads (ii) exactly seven heads (iii) atmost seven heads.

Solution:

X denote the number of heads appear

$$P(X = x) = {}^nC_x p^x q^{n-x}, x = 0, 1, 2, \dots, n$$

$$\text{Given: } p = P(\text{head}) = \frac{1}{2} \quad q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2} \text{ and } n = 10$$

$$\therefore X \sim B\left(10, \frac{1}{2}\right)$$

$$= {}^{10}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x}$$

$$P(X = x) = \frac{{}^{10}C_x}{1024}$$

(i) $P(\text{atleast seven heads})$

$$P(X \geq 7) = P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10)$$

$$= \frac{{}^{10}C_7}{1024} + \frac{{}^{10}C_8}{1024} + \frac{{}^{10}C_9}{1024} + \frac{{}^{10}C_{10}}{1024}$$

$$= \frac{120}{1024} + \frac{45}{1024} + \frac{10}{1024} + \frac{1}{1024}$$

$$= \frac{176}{1024}$$

(ii) $P(\text{exactly 7 heads})$

$$P(x=7) = \frac{{}^{10}C_7}{1024} = \frac{120}{1024}$$

(iii) $P(\text{atmost 7 heads})$

$$= P(x \leq 7) = 1 - P(x > 7)$$

$$= 1 - \{P(x = 8) + P(x = 9) + P(x = 10)\}$$

$$= 1 - 1 - \left\{ \frac{{}^{10}C_8}{1024} + \frac{{}^{10}C_9}{1024} + \frac{{}^{10}C_{10}}{1024} \right\}$$

$$= 1 - \frac{56}{1024}$$

$$= \frac{968}{1024}$$

Example 4

With usual notation find p for Binomial random variable X if $n = 6$ and $9 P(x = 4) = P(x = 2)$

Solution:

$$P(X = x) = {}^nC_x p^x q^{n-x}, x = 0, 1, 2, \dots, n$$

$$X \sim B(6, p) \Rightarrow P(X = x) = {}^6C_x p^x q^{6-x}$$

$$\text{Also } 9 \times P(X = 4) = P(X = 2)$$

$$\Rightarrow 9 \times {}^6C_4 p^4 q^2 = {}^6C_2 p^2 q^4$$

$$\Rightarrow 9 p^2 = q^2$$

$$\Rightarrow 3p = q \quad \text{as } p, q > 0$$

$$3p = 1 - p$$

$$4p = 1$$

$$\Rightarrow p = \frac{1}{4} = 0.25$$

Example 5

A Binomial distribution has parameters $n=5$ and $p=1/4$. Find the Skewness and Kurtosis.

Solution:

Here we are given $n=5$ and $p=\frac{1}{4}$

$$\begin{aligned}\text{Skewness} &= \frac{q-p}{\sqrt{npq}} \\ &= \frac{\frac{3}{4} - \frac{1}{4}}{\sqrt{5 \times \frac{1}{4} \times \frac{3}{4}}} \\ &= \frac{\frac{2}{4}}{\sqrt{\frac{15}{16}}} \\ &= \frac{2}{\sqrt{15}}\end{aligned}$$

Finding: The distribution is positively skewed.

Kurtosis

$$\begin{aligned}\text{Kurtosis} &= \frac{1-6pq}{npq} \\ &= \frac{1-6 \times \frac{1}{4} \times \frac{3}{4}}{\frac{15}{16}} \\ &= \frac{\frac{-2}{16}}{\frac{15}{16}} \\ &= \frac{-2}{15} \\ &= -0.1333\end{aligned}$$

Finding: The distribution is Platykurtic.

Example 6

In a Binomial distribution with 7 trials, $P(X=3)=P(X=4)$ Check whether it is a symmetrical distribution?

Solution:

A Binomial distribution is said to be symmetrical if $p = q = \frac{1}{2}$

Given: $P(X=3) = P(X=4)$

$$X \sim B(n, p)$$

$$P(X = x) = {}^nC_x p^x q^{n-x}, x = 0, 1, 2, \dots, n$$

$${}^nC_3 p^3 q^{n-3} = {}^nC_4 p^4 q^{n-4}$$

$${}^7C_3 p^3 q^4 = {}^7C_4 p^4 q^3 \quad \text{note that } {}^7C_3 = {}^7C_4$$

$$\text{On simplifying, we have } q = p$$

$$1-p = p$$

$$1 = 2p$$

$$p = \frac{1}{2}$$

$$q = \frac{1}{2}$$

Hence the given Binomial distribution is symmetrical.

Example 7

From a pack of 52 cards 4 cards are drawn one after another with replacement. Find the mean and variance of the distribution of the number of kings.

Solution:

Success X = event of getting king in a draw

p =probability of getting king in a single trial

$$p = \frac{4}{52}$$
$$= \frac{1}{13}$$

This is constant for each trial.

Hence, it is a binomial distribution with $n=4$ and $p = \frac{1}{13}$

$$\text{Mean} = np = 4 \times \frac{1}{13} = \frac{4}{13}$$

$$\text{Variance} = npq = \frac{4}{13} \times \frac{12}{13} = \frac{48}{169}$$

Example 8

In a street of 200 families, 40 families purchase the Hindu newspaper. Among the families a sample of 10 families is selected, find the probability that

- i. Only one family purchase the news paper
- ii. No family purchasing
- iii. Not more than one family purchase it

Solution:

$$X \sim B(n, p)$$

$$P(X = x) = {}^nC_x p^x q^{n-x}, x = 0, 1, 2, \dots, n$$

Let X denote the number of families purchasing Hindu Paper

p = Probability of their family purchasing the Hindu

$$p = \frac{40}{200} = \frac{1}{5}$$

$$q = \frac{4}{5}$$

$$n = 10$$

(i) Only one family purchase the Hindu

$$\begin{aligned} P(X=1) &= {}^nC_1 p^1 q^{n-1} \\ &= 10C_1 \times \frac{1}{5} \times \left(\frac{4}{5}\right)^9 \\ &= 2 \times \left(\frac{4}{5}\right)^9 \end{aligned}$$

(ii) No family purchasing the Hindu

$$\begin{aligned} P(X=0) &= {}^{10}C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10} \\ &= \left(\frac{4}{5}\right)^{10} \end{aligned}$$

(iii) Not more than one family purchasing The Hindu means that $X \leq 1$

$$\begin{aligned} P(X \leq 1) &= P[x=0] + P[x=1] \\ &= {}^{10}C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10} + {}^{10}C_1 \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^9 \\ &= \left(\frac{4}{5}\right)^{10} + 10 \times \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^9 \\ &= \left(\frac{4}{5}\right)^9 \left[\left(\frac{4}{5}\right) + 2 \right] \\ &= \left(\frac{4}{5}\right)^9 \left(\frac{14}{5}\right) \end{aligned}$$

Example 9

In a tourist spot, 80% of tourists are repeated visitors. Find the distribution of the numbers of repeated visitors among 4 selected peoples visiting the place. Also find its mode or the maximum visits by a visitor.

Solution:

Let the random variable X denote the number of repeated visitors.

$$X \sim B(n, p)$$

$$P(X = x) = {}^nC_x p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n$$

It is a Binomial Distribution with $n=4$

$$p = \frac{80}{100} = \frac{4}{5} \quad q = 1 - p = 1 - \frac{4}{5} = \frac{1}{5}$$

$$P(x=0) = {}^4C_0 \left(\frac{4}{5}\right)^0 \left(\frac{1}{5}\right)^4 = \left(\frac{1}{625}\right)$$

$$P(x=1) = {}^4C_1 \left(\frac{4}{5}\right)^1 \left(\frac{1}{5}\right)^3 = \frac{16}{625}$$

$$P(x=2) = {}^4C_2 \left(\frac{4}{5}\right)^2 \left(\frac{1}{5}\right)^2 = \left(\frac{96}{625}\right)$$

$$P(x=3) = {}^4C_3 \left(\frac{4}{5}\right)^3 \left(\frac{1}{5}\right)^1 = \left(\frac{256}{625}\right)$$

$$P(x=4) = {}^4C_4 \left(\frac{4}{5}\right)^4 \left(\frac{1}{5}\right)^0 = \left(\frac{256}{625}\right)$$

The probability distribution is given below.

$X=x$	0	1	2	3	4
$P(X=x)$	$\frac{1}{625}$	$\frac{16}{625}$	$\frac{96}{625}$	$\frac{256}{625}$	$\frac{256}{625}$

Example10

In a college, 60% of the students are boys. A sample of 4 students of the college, is taken, find the minimum number of boys should it have so that probability up to that number is $\geq 1/2$.

Solution:

It is given that 60% of the students of the college are boys and the selection probability for a boy is 60% or 0.6 As we are taking four samples, the number of trials $n = 4$. The selection process is independent.

$$X \sim B(n, p)$$

$$P(X = x) = {}^nC_x p^x q^{n-x}, x = 0, 1, 2, \dots, n$$

Let X be the number of boys so that $P(X \leq x) \geq \frac{1}{2}$

$$\text{If } x = 0 \quad P(X \leq 0) = {}^4C_0 \left(\frac{3}{5}\right)^0 \left(\frac{2}{5}\right)^4 = \frac{16}{625} < \frac{1}{2}$$

$$\begin{aligned} x=1 \quad P(X \leq 1) &= P(x=0) + P(x=1) \\ &= {}^4C_0 \left(\frac{3}{5}\right)^0 \left(\frac{2}{5}\right)^4 + {}^4C_1 \left(\frac{3}{5}\right)^1 \left(\frac{2}{5}\right)^3 \\ &= \frac{16}{625} + \frac{96}{625} = \frac{112}{625} < \frac{1}{2} \end{aligned}$$

$$\begin{aligned} x=2 \quad P(X \leq 2) &= P(x=0) + P(x=1) + P(x=2) \\ &= \frac{112}{625} + P(x=2) = \frac{112}{625} + {}^4C_2 \left(\frac{3}{5}\right)^2 \left(\frac{2}{5}\right)^2 \\ &= \frac{112}{625} + \frac{216}{625} = \frac{328}{625} > \frac{1}{2} \end{aligned}$$

Therefore the sample should contain a minimum of 2 boys.

POISSON DISTRIBUTION

Introduction

In a Binomial distribution with parameter n and p if the exact value of n is not definitely known and if p is very small then it is not possible to find the binomial probabilities. Even if n is known and it is very large, calculations are tedious. In such situations a distribution called Poisson distribution is very much useful.

In 1837 French mathematician Simeon Dennis Poisson derived the distribution as a limiting case of Binomial distribution. It is called after his name as Poisson distribution.

Conditions:

- i. The number of trials ' n ' is indefinitely large i.e., $n \rightarrow \infty$
- ii. The probability of a success ' p ' for each trial is very small i.e., $p \rightarrow 0$
- iii. $np = \lambda$ is finite
- iv. Events are Independent

Definition

A random variable X is said to follow a Poisson distribution if it assumes only non-negative integral values and its probability mass function is given by

$$P(X = x) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!}; & x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$



NOTE

- (i) λ is called the parameter of the Poisson distribution.
- (ii) $e = 1 + \frac{1}{1!} + \frac{1}{2!} + \dots = 2.71828\dots$ is an irrational number."

Characteristics of Poisson Distribution

- (i) Poisson distribution is a discrete distribution i.e., X can take values $0, 1, 2, \dots$
- (ii) p is small, q is large and n is indefinitely large i.e., $p \rightarrow 0$ $q \rightarrow 1$ and $n \rightarrow \infty$ and np is finite
- (iii) Values of constants : (a) Mean = λ = variance (b) Standard deviation = $\sqrt{\lambda}$ (c) Skewness = $1/\sqrt{\lambda}$ (iv) Kurtosis = $1/\lambda$
- (iv) It may have one or two modes
- (v) If X and Y are two independent Poisson variates, $X+Y$ is also a Poisson variate.
- (vi) If X and Y are two independent Poisson variates, $X-Y$ need not be a Poisson variate.
- (vii) Poisson distribution is positively skewed.
- (viii) It is leptokurtic.

Some examples:

- i. The event of a student getting first mark in all subjects and at all the examinations
- ii. The event of finding a defective item from the production of a reputed company
- iii. The number of blinds born in a particular year
- iv. The number of mistakes committed in a typed page
- v. The number of traffic accidents per day at a busy junction.
- vi. The number of death claims received per day by an insurance company.

Example 11

If 2% of electric bulbs manufactured by a certain company are defective find the probability that in a sample of 200 bulbs (i) less than 2 bulbs are defective (ii) more than 3 bulbs are defective.
[$e^{-4} = 0.0183$]

Solution:

Let X denote the number of defective bulbs

$$X \sim P(\lambda)$$

$$\therefore P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots, \infty$$

$$\text{Given } p = P(\text{a defective bulb}) = 2\% = \frac{2}{100} = 0.02$$

$$n = 200$$

$$\therefore \lambda = np = 200 \times 0.02 = 4$$

$$\therefore P(X = x) = \frac{e^{-4} 4^x}{x!}, \quad x = 0, 1, 2, \dots, \infty$$

(i) $P(\text{less than 2 bulbs are defective})$

$$= P(X < 2)$$

$$= P(x = 0) + P(x = 1)$$

$$= \frac{e^{-4} \cdot 4^0}{0!} + \frac{e^{-4} \cdot 4^1}{1!}$$

$$= e^{-4}(1 + 4)$$

$$= 0.0183 \times 5$$

$$= 0.0915$$

(ii) $P(\text{more than 3 defectives})$

$$= P(X > 3)$$

$$= 1 - P(X \leq 3)$$

$$= 1 - \{P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)\}$$

$$= 1 - \left\{ \frac{e^{-4} \cdot 4^0}{0!} + \frac{e^{-4} \cdot 4^1}{1!} + \frac{e^{-4} \cdot 4^2}{2!} + \frac{e^{-4} \cdot 4^3}{3!} \right\}$$

$$= 1 - e^{-4} \{1 + 4 + 8 + 10.667\}$$

$$= 1 - 0.0183 \times 23.667$$

$$= 0.567$$

Example 12

In a Poisson distribution $3P(X=2) = P(X=4)$. Find its parameter ' λ '

Solution:

The pmf of Poisson distribution is $P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}$, $x=0, 1, 2, \dots, \infty$,

Given $3P(X=2) = P(X=4)$

$$3 \cdot \frac{e^{-\lambda} \lambda^2}{2!} = \frac{e^{-\lambda} \lambda^4}{4!}$$

$$\lambda^2 = \frac{3 \times 4!}{2!} = 36$$

$$\therefore \lambda = 6 \text{ as } \lambda > 0$$

Example 13

Find the skewness and kurtosis of a Poisson variate with parameter 4.

Solution:

$$\lambda = 4$$

$$\text{Skewness} = \frac{1}{\sqrt{\lambda}} = \frac{1}{\sqrt{4}} = \frac{1}{2}$$

$$\text{Kurtosis} = \frac{1}{\lambda} = \frac{1}{4}$$

Example 14

If there are 400 errors in a book of 1000 pages, find the probability that a randomly chosen page from the book has exactly 3 errors.

Solution:

Let X denote the number of errors in pages

Let X denote the number of errors in pages

$$X \sim P(\lambda)$$

$$\therefore P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots, \infty$$

$$\text{The average number of errors per page} = \frac{400}{1000}$$

$$\text{i.e., } \lambda = \frac{400}{1000} = 0.4$$

$$\begin{aligned} P(X=3) &= \frac{e^{-\lambda} \lambda^3}{3!} = \frac{e^{-0.4} (0.4)^3}{3 \times 2 \times 1} \\ &= \frac{0.6703 \times 0.064}{6} \\ &= 0.00715 \end{aligned}$$

Example 15

If X is a Poisson variate with $P(X=0) = 0.2725$, find $P(X=1)$

Solution:

$$X \sim P(\lambda)$$

$$\therefore P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots, \infty$$

$$P(X=0) = 0.2725$$

$$\frac{e^{-\lambda} \lambda^0}{0!} = 0.2725$$

$$e^{-\lambda} = 0.2725$$

$$\lambda = 1.3 \text{ (from the table of values of } e^{-m} \text{)}$$

$$\begin{aligned} P(x=1) &= \frac{e^{-\lambda} \lambda^1}{1!} = \frac{e^{-1.3} \times 1.3^1}{1!} \\ &= 0.2725 \times 1.3 \\ &= 0.3543 \end{aligned}$$

Example 16

The probability of safety pin manufactured by a firm to be defective is 0.04. (i) Find the probability that a box containing 100 such pins has one defective pin. (ii) Among 200 such boxes, how many boxes will have no defective pin.

Solution:

Let X denote the number boxes with defective pins

$$X \sim P(\lambda)$$

$$\therefore P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots, \infty$$

$$p = 0.04$$

$$n = 100$$

$$\lambda = np = 4$$

$$\begin{aligned} \text{(i)} \quad P(X=1) &= \frac{e^{-\lambda} \lambda}{1!} = e^{-4}(4) = 0.0183 \times 4 \\ &= 0.0732 \end{aligned}$$

$$\text{(ii)} \quad P(X=0) = \frac{e^{-\lambda} \lambda^0}{0!} = e^{-4} = 0.0183$$

$$\begin{aligned} \text{Number of boxes having no defective pin} &= 200 \times 0.0183 \\ &= 3.660 \\ &= 4 \end{aligned}$$

Continuous distributions:

Rectangular or Uniform Distribution

Definition

A random variable X is said to have a continuous Uniform distribution over the interval (a, b) if its probability density function is

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Characteristics of Uniform Distribution

- i. a and b are the parameters of the Uniform distribution and we write $X \sim U(a, b)$
- ii. The distribution is also known as Rectangular distribution, as the curve
- iii. $y = f(x)$ describes a rectangle over the x-axis and between ordinates at $x = a$ and $x = b$.
- (iv) $X \sim U(-a, a)$ then its p.d.f. is

$$f(x) = \begin{cases} \frac{1}{2a} & -a < x < a \\ 0 & \text{otherwise} \end{cases}$$

Constants of uniform distribution $X \sim U(a, b)$

- (i) Mean $\mu = \frac{a+b}{2}$
- (ii) Variance $\sigma^2 = \frac{(b-a)^2}{12}$
- (iii) Median $= \frac{a+b}{2}$
- (iv) Skewness $= 0$
- (v) Kurtosis $= -\frac{6}{5}$
- (vi) $Q_1 = \frac{3a+b}{4}$
- (vii) $Q_3 = \frac{a+3b}{4}$

Example 17

If $X \sim U(200, 250)$ find its p.d.f and $P(X > 230)$

Solution:

(i) For $X \sim U(a, b)$

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Taking $a = 200$ and $b = 250$ the required p.d.f. is

$$\begin{aligned} f(x) &= \begin{cases} \frac{1}{250-200} & 200 < x < 250 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{50} & 200 < x < 250 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

$$(ii) \quad P(x > 230) = \int_{230}^{\infty} f(x) dx = \int_{230}^{250} \frac{1}{50} dx$$

$$= \frac{1}{50} [x]_{230}^{250} = \frac{250-230}{50} = \frac{20}{50} = 0.4$$

Example 18

If X is a Uniform variate with the parameter 50 and 100, find the mean, median and standard deviation.

Solution:

$$X \sim U(50, 100)$$

$$\text{Here } a=50 \quad b=100$$

$$\text{mean} = \text{median} = \frac{a+b}{2} = \frac{150}{2} = 75$$

$$\begin{aligned} \text{S.D} &= \sqrt{\frac{(b-a)^2}{12}} = \frac{b-a}{\sqrt{12}} = \frac{100-50}{\sqrt{12}} \\ &= \frac{50}{\sqrt{12}} = \frac{50}{3.464} \\ &= 14.434 \end{aligned}$$

Example 19

If X is a random variable having a uniform distribution $U(a, b)$ such that $P(20 < X < 40) = 0.2$ and mean = 150, find a and b .

For $X \sim U(a, b)$

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

$$P(20 < X < 40) = 0.2$$

$$\int_{20}^{40} \frac{1}{b-a} dx = 0.2$$

$$\frac{1}{b-a} (x)_{20}^{40} = 0.2$$

$$\frac{1}{b-a} (20) = 0.2$$

$$b - a = 100 \quad \dots (1)$$

$$\text{mean} = 150$$

$$\frac{a+b}{2} = 150$$

$$a+b = 300 \quad \dots (2)$$

(1)+(2) implies $2b = 400$ and $b = 200$.

Substituting b in (2) we have $a + 200 = 300$ and that $a = 100$.

$a=100, b=200$

$X \sim U(100, 200)$

Example 20

If X is a Uniform variable $U(a,b)$ with first and third quartiles 100 and 200, find the p.d.f of X .

Solution:

$$Q_1 = 100$$

$$\frac{3a+b}{4} = 100$$

$$3a+b = 400 \quad \dots (1)$$

$$Q_3 = 200$$

$$\frac{a+3b}{4} = 200$$

$$a+3b = 800 \quad \dots (2)$$

Solving (1) and (2) we get

$$f(x) = \begin{cases} \frac{1}{b-a} & a=50, b=250 \\ 0 & \text{otherwise} \end{cases} = \frac{1}{250-50} = \frac{1}{200}, 50 < x < 250$$

Example 21

Electric trains on a certain line run every 15 minutes between mid- night and six in the morning. What is the probability that a man entering the station at a random time during this period will have to wait at least ten minutes?

Let the random variable X denote the waiting time (in minutes).

The given assumption indicates that X is distributed Uniformly on $(0,15)$.

$$\begin{aligned} P(X > 10) &= \int_{10}^{15} f(x) dx = \frac{1}{15} \int_{10}^{15} 1 dx \\ &= \frac{1}{15} (x)_{10}^{15} = \frac{1}{15} (15 - 10) = \frac{1}{3} \end{aligned}$$

Continuous distributions:

NORMAL DISTRIBUTION

Introduction

The Normal distribution was first discovered by the English Mathematician De-Moivre in 1733, and he obtained this distribution as a limiting case of Binomial distribution and applied it to problems arising in the game of chance. In 1774 Laplace used it to estimate historical errors and in 1809 it was used by Gauss as the distribution of errors in Astronomy. Thus throughout the 18th and 19th centuries efforts were made for a common law for all continuous distributions which was then known as the Normal distribution.

Definition

A random variable X is said to have a Normal distribution with parameters μ and σ^2 if its probability density function is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2}$ where
 $-\infty < x < \infty$, $-\infty < \mu < \infty$ and $\sigma > 0$

It is denoted by $X \sim N(\mu, \sigma^2)$

Here μ is called as mean and σ^2 is variance of the distribution.



NOTE

(i) $B(X: n, p)$ when $n \rightarrow \infty$ and p, q are not small will become a Normal distribution.

(i) $P(X, \lambda)$ when $\lambda \rightarrow \infty$ will become a Normal distribution.

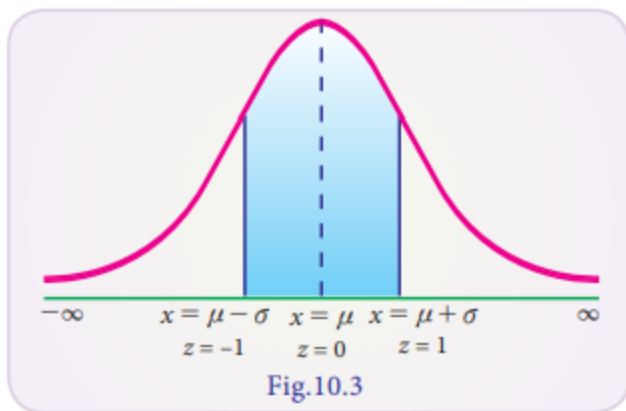
(ii) When $X \sim N(\mu, \sigma^2)$ then $Z = \frac{x - \mu}{\sigma}$ then $Z \sim N(0, 1)$

$$\text{i.e., } f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad -\infty \leq z \leq \infty$$

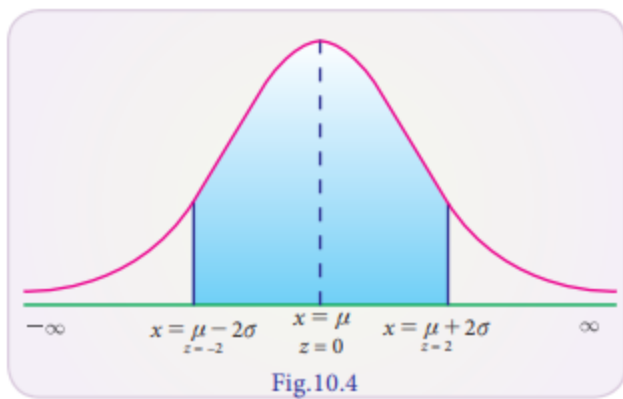
Z is known as a Standard Normal variate with mean 0 and variance 1.

To find probabilities of X we convert X into Z and then make use of standard normal table.

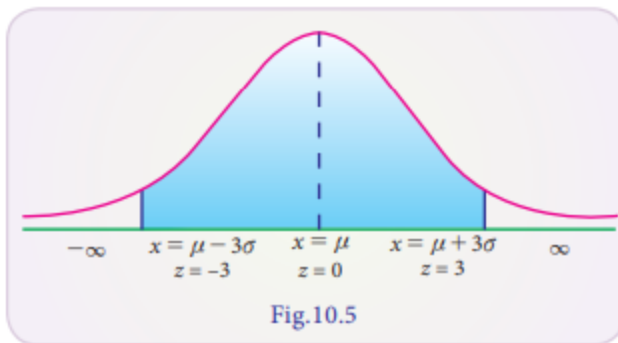
$$\begin{aligned}
 \text{(c) } P(\mu - \sigma < X < \mu + \sigma) &= P(-1 < Z < 1) \\
 &= 2 P(0 < Z < 1) \\
 &= 2 (0.3413) \\
 &= 0.6826
 \end{aligned}$$



$$\begin{aligned}
 \text{(d) } P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(-2 < Z < 2) \\
 &= 2 P(0 < Z < 2) \\
 &= 2 (0.4772) \\
 &= 0.9544
 \end{aligned}$$



$$\begin{aligned}
 (e) \quad & P(\mu - 3\sigma < X < \mu + 3\sigma) \\
 &= P(-3 < Z < 3) \\
 &= 2 P(0 < Z < 3) \\
 &= 2 (0.49865) \\
 &= 0.9973
 \end{aligned}$$



$$\begin{aligned}
 (f) \quad & P(|X - \mu| > 3\sigma) = P(|Z| > 3) \\
 &= 1 - P(|Z| < 3) \\
 &= 1 - P(-3 < Z < 3) \\
 &= 1 - 0.9973
 \end{aligned}$$

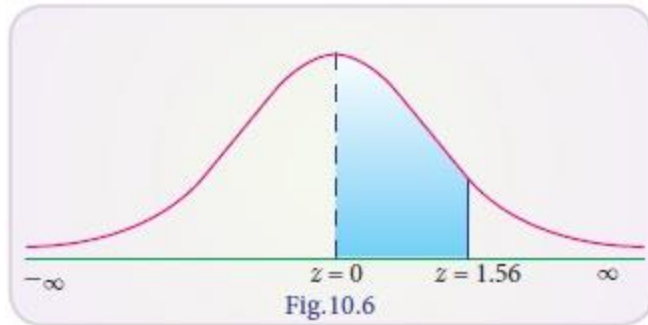
Example 22

Find the area between $z = 0$ and $z = 1.56$

Solution:

$$P(0 < Z < 1.56) = 0.4406$$

(from the Normal Probability table)



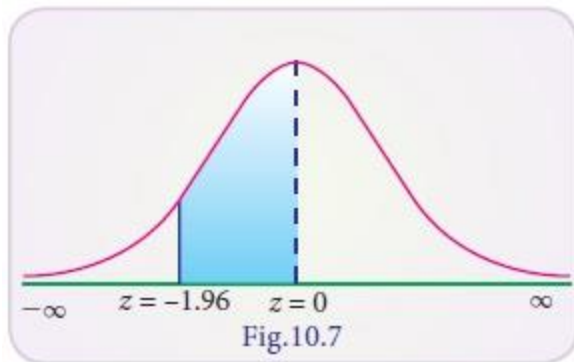
Example 23

Find the area of the Standard Normal variate from -1.96 to 0

Solution:

$$P(-1.96 < Z < 0) = P(0 < Z < 1.96) \text{ (by symmetry)}$$

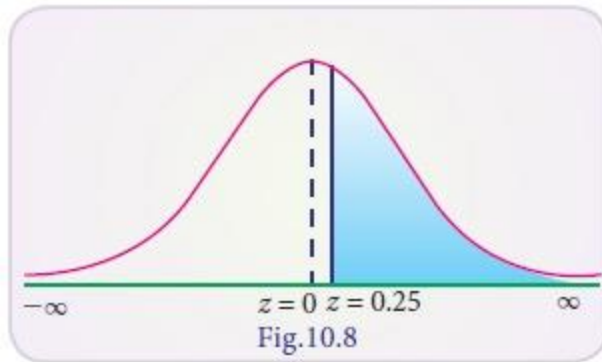
$$= 0.4750 \text{ (from the Normal Probability table)}$$



Example 24

Find the area to the right of $Z = 0.25$

Solution:

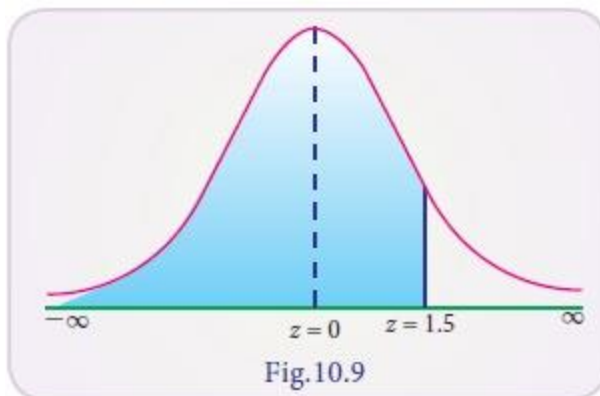


$$\begin{aligned} P(Z > 0.25) &= P(0 < Z < \infty) - P(0 < Z < 0.25) \\ &= 0.5000 - 0.0987 \text{ (from table)} \\ &= 0.4013 \end{aligned}$$

Example 25

Find the area to the left of $Z = 1.5$

Solution:



$$\begin{aligned} P(Z < 1.5) &= P(-\infty < Z < 0) + P(0 < Z < 1.5) \\ &= 0.5000 + 0.4332 \text{ (from table)} \end{aligned}$$

$$= 0.9332$$

Example 26

Find the area between $Z = -1$ and $Z = 1.75$

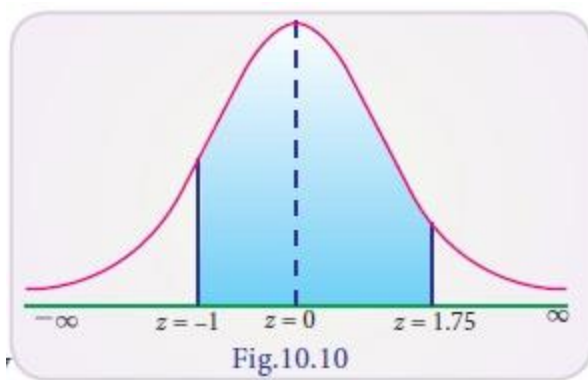
Solution:

$$P(-1 < Z < 1.75)$$

$$P(-1 < Z < 0) + P(0 < Z < 1.75)$$

$$= P(0 < Z < 1) + P(0 < Z < 1.75) \text{ by symmetry}$$

$$= 0.3413 + 0.4599 = 0.8012$$



Example 27

Find the maximum value of the p.d.f of the Normal distribution with mean 40 and standard deviation 10. Also find its points of inflection.

Solution:

$$\mu = 40$$

$$\sigma = 10$$

$$\text{Maximum value} = f(x)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} = \frac{1}{10\sqrt{2\pi}}$$

Points of inflection = $\mu \pm \sigma$

$$= 40 \pm 10$$

$$= 30 \text{ and } 50$$

Example 28

A Normal variable X has the mean 50 and the S.D 5. Find its mean deviation about mean and the quartile deviation.

Solution:

Given Mean $= \mu = 50$, standard deviation $\sigma = 5$

Mean deviation above mean $= 0.8 \sigma$

$$= 0.8 \times 5 = 4$$

Quartile deviation $= 0.6745\sigma$

$$= 0.6745 \times 5$$

$$= 3.3725$$

Example 29

Find the quartiles of the Normal distribution having mean 60 and S.D 10.

Solution:

$$\mu = 60, \sigma = 10$$

Let x_1 be the value such that the area from x_1 to μ is 25%

$$P(x_1 < X < \mu) = 25\% = 0.25$$

$$P(z_1 < Z < 0) = 0.25 \text{ where } z_1 = \frac{x_1 - \mu}{\sigma} = \frac{x_1 - 60}{10}$$

From Normal Probability table

$$P(-0.675 < Z < 0) = 0.25015$$

$$\therefore z_1 = -0.675$$

$$\begin{aligned} \therefore \frac{x_1 - \mu}{\sigma} &= -0.675 \\ \frac{x_1 - 60}{10} &= -0.675 \\ X_1 - 60 &= -6.75 \end{aligned}$$

$$X_1 = 53.25 \text{ that is } Q_1 = 53.25$$

Let x_2 be the value, so that

$$\therefore P(\mu < X < x_2) = 0.25$$

$$P(0 < Z < z_2) = 0.25, \text{ where } z_2 = \frac{x_2 - \mu}{\sigma} = \frac{x_2 - 60}{10}$$

From the table

$$P(0 < Z < 0.675) = 0.25015$$

$$\therefore Z_2 = 0.675$$

$$\frac{x_2 - 60}{10} = 0.675$$

$$X_2 = 66.75 \quad \text{that is } Q_2 = 66.75$$

Example 30

The height of the rose plants in a garden is Normally distributed with a mean 100cms. Given that 10% of the plants have height greater than 104cm. Find (i) The S.D of the distribution (ii) The number of plants have height greater than 105cms if there were 200 plants in the garden.

Solution:

The S.D of the distribution

Let X be the height of the rose plants that is Normally distributed.

Given: $P(100 < x < 104) = 40\%$

$$P(0 < Z < 4/\sigma) = 0.40 \quad \dots (1)$$

But from table of area under Standard Normal curve

$$P(0 < Z < 1.28) = 0.3997 \quad \dots (2)$$

From (1) and (2)

$$4/\sigma = 1.28$$

$$\sigma = 3.125$$

(ii) The number of plants with height greater than 105 cms if there were 200 plants in the garden.

$$X = 105$$

$$\begin{aligned} Z &= \frac{x - \mu}{\sigma} = \frac{105 - 100}{3.125} \\ &= \frac{5}{3.125} \end{aligned}$$

$$= 1.6$$

$$P(1.6 < Z < \infty) = 0.5 - 0.4452 = 0.0548$$

$$\text{Number of plants} = 200 \times 0.0548 = 10.9600 \approx 11$$

∴ 11 plants have height greater than 105cms.

Example 31

Students of a class were given an aptitude test. Their marks were found to be normally distributed with mean 60 and standard deviation 5. What percentage of students scored (i) more than 60 marks (ii) less than 56 marks (iii) between 45 and 65 marks.

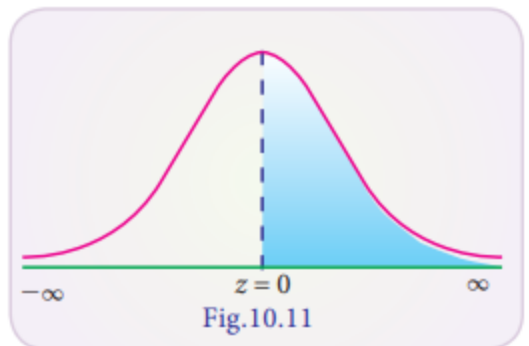
Solution:

Given mean $\mu = 60$ and standard deviation $\sigma = 5$

Standard normal variate $Z = \frac{x - \mu}{\sigma} = \frac{x - 60}{5}$

$$\begin{aligned} \text{(i) } P(\text{more than } 60) &= P(x > 60) \\ &= P\left(z > \frac{60 - 60}{5}\right) \\ &= P(Z > 0) \\ &= P(0 < Z < \infty) \\ &= 0.5000 \end{aligned}$$

Student scored more than 60 marks = 0.5000×100
 $= 50\%$



$$(ii) P(\text{less than 56 marks}) = P(X < 56)$$

$$= P\left(Z < \frac{56 - 60}{5}\right)$$

$$= P(Z < -0.8)$$

$$= P(-\infty < Z < 0) - P(-0.8 < Z < 0)$$

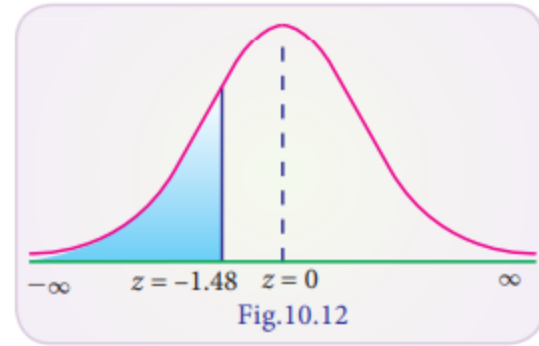
$$= 0.5000 - P(0 < Z < 0.8)$$

$$= 0.5000 - 0.2881$$

$$= 0.2119$$

$$\therefore \text{Number of students scored less than 50 marks} = 0.2119 \times 100$$

$$= 21.19\%$$



$$(iii) P(\text{between 45 and 65 marks}) = P(45 < x < 65)$$

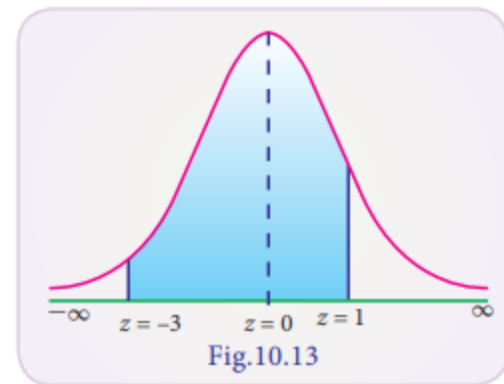
$$= P\left(\frac{45 - 60}{5} < z < \frac{65 - 60}{5}\right)$$

$$= P(-3 < z < 1)$$

$$= P(0 < Z < 3) + P(0 < Z < 1)$$

$$= 0.4986 + 0.3413$$

$$= 0.8399$$



$$\text{Number of students scored between 45 and 60 marks} = 0.8399 \times 100 = 83.99\%$$

$$\text{Number of students scored between 45 and 60 marks} = 0.8399 \times 100 = 83.99\%$$

Example 32

X has Normal distribution with mean 2 and standard deviation 3. Find the value of the variable x such that the probability of the interval from the mean to that value is 0.4115.

Solution:

$$\text{Given } \mu = 2, \sigma = 3, z = \frac{x - \mu}{\sigma} = \frac{x - 2}{3}$$

$$\text{Let } Z_1 = \frac{x_1 - 2}{3}$$

$$\text{We have } P(\mu < x < x_1) = 0.4115$$

$$P(0 < Z < z_1) = 0.4115$$

$$\text{But } P(0 < Z < 1.35) = 0.4115 \text{ (from the Normal Probability table)}$$

$$\therefore Z_1 = 1.35 \therefore \frac{x_1 - 2}{3} = 1.35 \text{ (or) } x_1 = (1.35) \times 3 + 2 = 6.05$$

Example 33

In a Normal distribution 7% items are under 35 and 89% are under 63. Find its mean and standard deviation.

Solution:

$$Z = \frac{x - \mu}{\sigma}$$

We have $P(x < 35) = 7\% = 0.07$

If $z_1 =$ then $P(z_1 < Z < 0) = 0.50 - 0.07 = 0.43$

$\therefore z_1 = -1.48$ (from Normal Probability table)

$$\frac{35 - \mu}{\sigma} = -1.48$$

$$35 - \mu = -1.48\sigma \quad \dots (1)$$

Also $P(X < 63) = 89\% = 0.89$

If $Z_2 =$ then $P(0 < Z < Z_2) = 0.89 - 0.50 = 0.39$

$Z_2 = 1.23$ (from Normal Probability table)

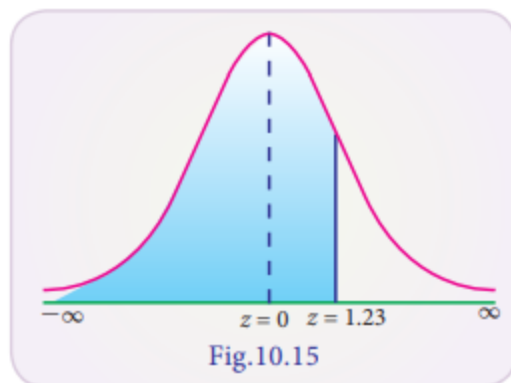
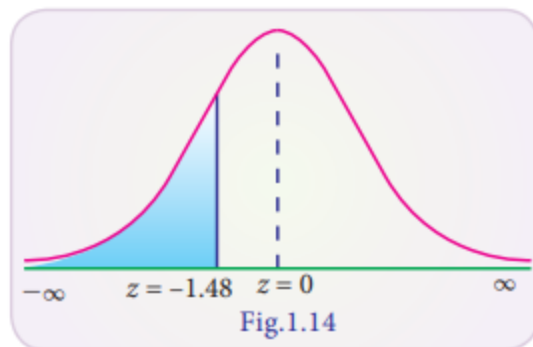
$$\frac{63 - \mu}{\sigma} = 1.23 \text{ (or) } 63 - \mu = 1.23\sigma \quad \dots (2)$$

$$(2) - (1) \Rightarrow 28 = 2.71\sigma \text{ or } \sigma = \frac{28}{2.71} = 10.33$$

From (1) we have

$$\mu = 35 + 1.48\sigma = 35 + 1.48(10.33)$$

$$\mu = 50.27$$





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

QUESTION BANK

UNIT IV

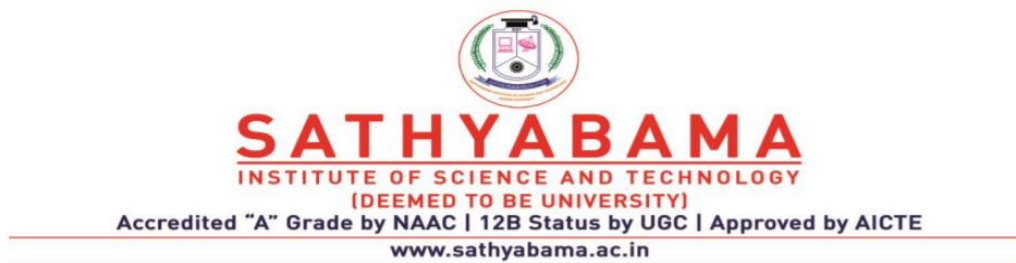
PART A

1. Define binomial distribution
2. Write the formula for Mean and Variance of binomial distribution
3. State the formula for binomial distribution.
4. Lists the formula for Mean and Variance of Poisson distribution.
5. Derive recurrence relation for the Poisson distribution.
6. Write the properties of normal distribution
7. Define Uniform distribution.
8. State the formula for exponential distribution.
9. State memory less property of exponential distribution.
10. State Geometric distribution.

PART B

1. In a large consignment of electric bulbs 10% are defective. A random sample of 20 is taken for inspection. Analyze the probability that (i) all are good bulbs (ii) atmost there are 3 defective bulbs and exactly there are three defective bulbs.
2. In certain factory turning razor blades there is a small chance of $1/500$ for any blade to be defective. The blades are in packets of 10. Use Poisson distribution to solve the approximate number of packets containing two defectives in a consignment of 10000 packets.
3. Ten coins are thrown simultaneously. Find the probability of getting at least 7 heads.
4. Six coins are tossed 6400 times. Using the poisson distribution, what is the approximate probability of getting six heads 10 times.
5. Find the probability that atmost 5 defective fuses will be found in a box of 200 fuses if experiences shows that 2% of such fuses are defective.
6. Subway trains on certain line run every half hour between mid night and six in the morning. What is the probability that a man entering the station at a random time during this period will have to wait at least twenty minutes.
7. The weekly wages of 1000 workmen are normally distributed around a mean of Rs.70 with a S.D of Rs.5. Estimate the number of workers whose weekly wages will be (i) between Rs.69 and Rs.72 (ii) less than Rs.69, (iii) More than Rs.72.
8. The time (in hours) required to repair a machine is exponentially distributed with parameter $\lambda=1/2$.
(a) What is the probability that the repair time exceeds $2n$?
(b) What is the conditional probability that a repair takes at 11h given that its direction exceeds 8h?
9. The mileage which car owners get with certain kind of radial tyre is a RV having an exponential distribution with mean 4,000 km. Find the Probabilities that one of these tires will last (i) at least 2000 km (ii) at most 3000 km.
10. Fit a Poisson distribution to the following data and calculate theoretical Frequencies.

Deaths	0	1	2	3	4
Frequencies	122	60	15	2	1



UNIT V

TIME SERIES

TIME SERIES

G. E. P. Box (1919-2013) was a British Statistician was “one of the great statistical minds” of the 20th century, who received his Ph.D., from the University of London, under the supervision of E. S. Pearson. He served as President of Americal Statistical Association in 1978 and of the Institute of Mathematics in 1979. His name is associated with Box-Cox transformation in addition to Box-Jenkins models in time series.

Introduction

In modern times we see data all around. The urge to evaluate the past and to peep into the future has made the need for forecasting. There are many factors which change with the passage of time. Sometimes sets of observations which vary with the passage of time and whose measurements made at equidistant points may be regarded as time series data. Statistical data which are collected, observed or recorded at successive intervals of time constitute time series data. In the study of time series, comparison of the past and the present data is made. It also compares two or more series at a time. The purpose of time series is to measure chronological variations in the observed data.

In an ever changing business and economic environment, it is necessary to have an idea about the probable future course of events. Analysis of relevant time series helps to achieve this, especially by facilitating future business forecasts. Such forecasts may serve as crucial inputs in deciding competitive strategies and planning growth initiatives.

DEFINITION

Time series refers to any group of statistical information collected at regular intervals of time. Time series analysis is used to detect the changes in patterns in these collected data.

1. Definition by Authors

According to Mooris Hamburg “A time series is a set of statistical observations arranged in chronological order”.

Ya-Lun-Chou : “A time series may be defined as a collection of readings belonging to different time periods of some economic variable or composite of variables”.

W.Z. Hirsch says “The main objective in analyzing time series is to understand, interpret and evaluate change in economic phenomena in the hope of more correctly anticipating the course of future events”.

Series 2. Uses of Time

- ☐ Time series is used to predict future values based on previously observed values.
- ☐ Time series analysis is used to identify the fluctuation in economics and business.
- ☐ It helps in the evaluation of current achievements.
- ☐ Time series is used in pattern recognition, signal processing, weather forecasting and earthquake prediction.

It can be said that time series analysis is a big tool in the hands of business executives to plan their sales, prices, policies and production.

COMPONENTS OF TIME SERIES

The factors that are responsible for bringing about changes in a time series are called the components of time series.

Components of Time Series

1. Secular trend
2. Seasonal variation
3. Cyclical variation
4. Irregular (random) variation

Approaches to time series

There are two approaches to the decomposition of time series data

- (i) Additive approach
- (ii) Multiplicative approach

The above two approaches are used in decomposition, depending on the nature of relationship among the four components.

1. The additive approach

The additive approach is used when the four components of a time series are visualized as independent of one another. Independence implies that the magnitude and pattern of movement of the components do not affect one another. Under this assumption the magnitudes of the time series are regarded as the sum of separate influences of its four components.

$$Y = T + C + S + R$$

where Y = magnitude of a time series

T = Trend,

C = Cyclical component,

S = Seasonal component, and

R = Random component

In additive approach, the unit of measurements remains the same for all the four components.

2. The Multiplicative approach

The multiplicative approach is used where the forces giving rise to the four types of variations are visualized as interdependent. Under this assumption, the magnitude of the time series is the product of its four components.

$$i.e. Y = T \times C \times S \times R$$

Difference between the two approaches

Multiplicative	Additive
(i) Four components of time series are interdependent	Four components of time series are independent
(ii) Logarithm of components are additive	Components are additive

Multiplicative

1. Four components of time series are interdependent
2. Logarithm of components are additive

Additive

1. Four components of time series are independent
2. Components are additive

Measurements of Components

(i) Secular trend (ii) Seasonal variation (iii) Cyclical variation (iv) Irregular variation

i) Secular trend

It refers to the long term tendency of the data to move in an upward or downward direction. For example, changes in productivity, increase in the rate of capital formation, growth of population, *etc.*., follow secular trend which has upward direction, while deaths due to improved medical facilities and sanitations show downward trend. All these forces occur in slow process and influence the time series variable in a gradual manner.

Methods of Measuring Trend

Trend is measured using by the following methods:

1. Graphical method
2. Semi averages method
3. Moving averages method
4. Method of least squares

(ii) Seasonal variation

Seasonal variations are fluctuations within a year over different seasons.

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

Methods of constructing seasonal indices

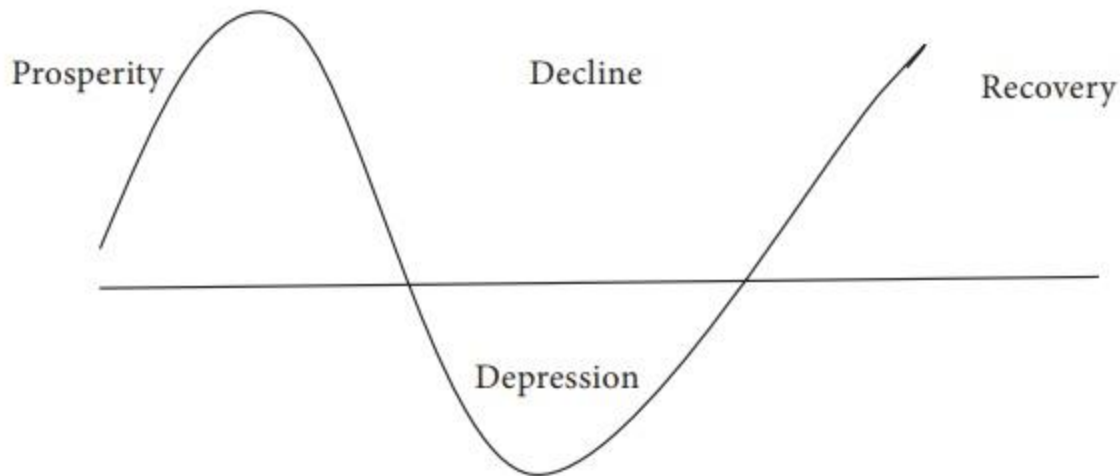
There are four methods of constructing seasonal indices.

1. Simple averages method
2. Ratio to trend method
3. Percentage moving average method
4. Link relatives method

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

(iii) Cyclical variation

Cyclical variations refer to periodic movements in the time series about the trend line, described by upswings and downswings. They occur in a cyclical fashion over an extended period of time (more than a year). For example, the business cycle may be described as follows.



The cyclical pattern of any time series tells about the prosperity and recession, ups and downs, booms and depression of a business. In most of the businesses there are upward trend for some time followed by a downfall, touching its lowest level. Again a rise starts which touches its peak. This process of prosperity and recession continues and may be considered as a natural phenomenon.

(iv) Irregular variation

In practice, the changes in a time series that cannot be attributed to the influence of cyclic fluctuations or seasonal variations or those of the secular trend are classified as irregular variations.

In the words of Patterson, “Irregular variation in a time series is composed of non-recurring sporadic (rare) form which is not attributed to trend, cyclical or seasonal factors”.

Nothing can be predicted about the occurrence of irregular influences and the magnitude of such effects. Hence, no standard method has been evolved to estimate the same. It is taken as the residual left in the time series, after accounting for the trend, seasonal and cyclic variations.

Secular trend

It refers to the long term tendency of the data to move in an upward or downward direction. For example, changes in productivity, increase in the rate of capital formation, growth of population, *etc.* , follow secular trend which has upward direction, while deaths due to improved

medical facilities and sanitations show downward trend. All these forces occur in slow process and influence the time series variable in a gradual manner.

Methods of Measuring Trend

Trend is measured using by the following methods:

1. Graphical method
2. Semi averages method
3. Moving averages method
4. Method of least squares

Method 1. Graphical

Under this method the values of a time series are plotted on a graph paper by taking time variable on the X-axis and the values variable on the Y-axis. After this, a smooth curve is drawn with free hand through the plotted points. The trend line drawn above can be extended to forecast the values. The following points must be kept in mind in drawing the freehand smooth curve.

- (i) The curve should be smooth
- (ii) The number of points above the line or curve should be approximately equal to the points below it
- (iii) The sum of the squares of the vertical deviation of the points above the smoothed line is equal to the sum of the squares of the vertical deviation of the points below the line.

Merits

- ☐ It is simple method of estimating trend.
- ☐ It requires no mathematical calculations.
- ☐ This method can be used even if trend is not linear.

Demerits

- ☐ It is a subjective method

□ The values of trend obtained by different statisticians would be different and hence not reliable.

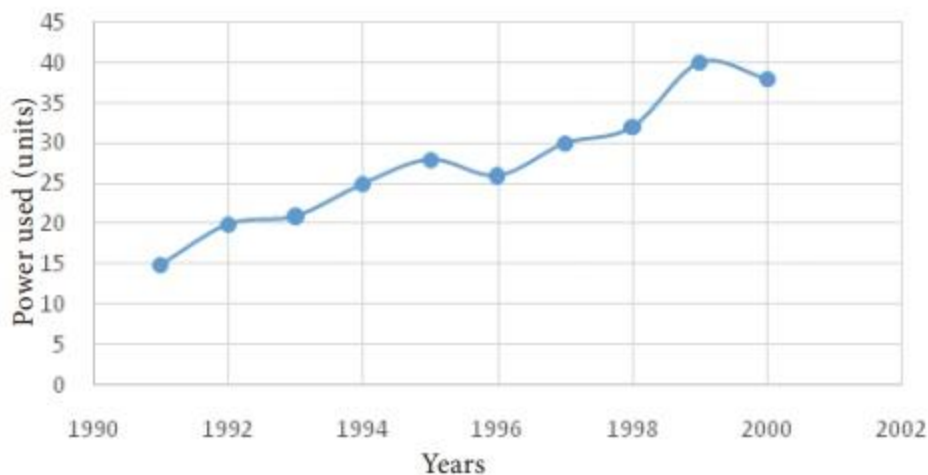
Example 1

Annual power consumption per household in a certain locality was reported below.

Years	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Power used (units)	15	20	21	25	28	26	30	32	40	38

Draw a free hand curve for the above data.

Solution:



Method 2. Semi-Average

In this method, the series is divided into two equal parts and the average of each part is plotted at the mid-point of their time duration.

(i) In case the series consists of an even number of years, the series is divisible into two halves. Find the average of the two parts of the series and place these values in the mid-year of each of the respective durations.

(ii) In case the series consists of odd number of years, it is not possible to divide the series into two equal halves. The middle year will be omitted. After dividing the data into two parts, find the arithmetic mean of each part. Thus we get semi-averages.

(iii) The trend values for other years can be computed by successive addition or subtraction for each year ahead or behind any year.

Merits

- ☐ This method is very simple and easy to understand
- ☐ It does not require many calculations.

Demerits

- ☐ This method is used only when the trend is linear.
- ☐ It is used for calculation of averages and they are affected by extreme values.

Example 2

Calculate the trend values using semi-averages methods for the income from the forest department. Find the yearly increase.

Year	2008	2009	2010	2011	2012	2013
Income (in crores)	46.17	51.65	63.81	70.99	84.91	91.64

Source: The Principal Chief conservator of forests, Chennai-15.

Solution:

Year	Income	3-Year semi-total	Semi-average
2008	46.17	161.63	53.877
2009	51.65		
2010	63.81		
2011	70.99	247.54	82.513
2012	84.91		
2013	91.64		

Difference between the central years = $2012 - 2009 = 3$

Difference between the semi-averages = $82.513 - 53.877 = 28.636$

Increase in trend value for one year = $28.636 / 3 = 9.545$

Trend values for the previous and successive years of the central years can be calculated by subtracting and adding respectively, the increase in annual trend value.

Example 3

Population of India for 7 successive census years are given below. Find the trend values using semi-averages method.

Census Year	1951	1961	1971	1981	1991	2001	2011
Population (in lakhs)	301.2	336.9	412.0	484.1	558.6	624.1	721.4

Solution:

Trend values using semi average method

Census Year	Population (in lakhs)	3-year semi-total	3-year semi-average	Trend values
1951	301.2	1050.1	350.03	278.86
1961	336.9			350.03
1971	412.0			421.2
1981	484.1			492.37
1991	558.6	1904.1	634.7	563.54
2001	624.1			634.71
2011	721.4			705.88

Difference between the years = $2001 - 1961 = 40$

Difference between the semi-averages = $634.7 - 350.03 = 284.67$

Increase in trend value for 10 year = $284.67 / 4 = 71.17$

For example the trend value for the year 1951 = $350.03 - 71.17 = 278.86$ The value for the year 2011 = $634.7 + 71.17 = 705.87$

The trend values have been calculated by successively subtracting and adding the increase in trend for previous and following years respectively.

Example 4

Find the trend values by semi-average method for the following data.

Year	1965	1966	1967	1968	1969	1970	1971	1972
Production of bleaching powder (in tonnes)	7.4	10.8	9.2	10.5	15.5	13.7	16.7	15

Solution:

Trend values using semi averages method

Year	Production of bleaching powder	4 year semi-total	4 year semi-average	Trend
1965	7.4	37.9	9.475	7.315
1966	10.8			8.755
1967	9.2			10.195
1968	10.5			11.635
1969	15.5	60.9	15.225	13.075
1970	13.7			14.515
1971	16.7			15.955
1972	15			17.395

Difference between the years = $1970.5 - 1966.5 = 4$

Difference between the semi-averages = $15.225 - 9.475 = 5.75$

Increase in trend = $5.75 / 4 = 1.44$

Half yearly increase in trend = $1.44 / 2 = 0.72$

The trend value for 1967 = $9.475 + 0.72 = 10.195$

The trend value for 1968 = $9.475 + 3 * 0.72 = 11.635$

Similarly the trend values for the other years can be calculated.

Method Averages 3. Moving

Moving averages is a series of arithmetic means of variate values of a sequence. This is another way of drawing a smooth curve for a time series data.

Moving averages is more frequently used for eliminating the seasonal variations. Even when applied for estimating trend values, the moving average method helps to establish a trend line by eliminating the cyclical, seasonal and random variations present in the time series. The period of the moving average depends upon the length of the time series data.

The choice of the length of a moving average is an important decision in using this method.

For a moving average, appropriate length plays a significant role in smoothening the variations.

In general, if the number of years for the moving average is more then the curve becomes smooth.

Merits

- ☐ It can be easily applied
- ☐ It is useful in case of series with periodic fluctuations.
- ☐ It does not show different results when used by different persons
- ☐ It can be used to find the figures on either extremes; that is, for the past and future years.

Demerits

- ☐ In non-periodic data this method is less effective.
- ☐ Selection of proper 'period' or 'time interval' for computing moving average is difficult.
- ☐ Values for the first few years and as well as for the last few years cannot be found.

Moving averages odd number of years (3 years)

To find the trend values by the method of three yearly moving averages, the following steps have to be considered.

- ☐ Add up the values of the first 3 years and place the yearly sum against the median year. [This sum is called moving total]
- ☐ Leave the first year value, add up the values of the next three years and place it against its median year.
- ☐ This process must be continued till all the values of the data are taken for calculation.
- ☐ Each 3-yearly moving total must be divided by 3 to get the 3-year moving averages, which is our required trend values.

Example 5

Calculate the 3-year moving averages for the loans issued by co-operative banks for non-farm sector/small scale industries based on the values given below.

Year	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15
Loan by District Central Cooperative banks (Rupees in crores)	41.82	40.05	39.12	24.72	26.69	59.66	23.65	28.36	33.31	31.60	36.48

Solution:

The three year moving averages are shown in the last column.

Year	Loan by District Central Cooperative Banks	3-year moving total	3-year moving average
2004-05	41.82	-	-
2005-06	40.05	120.99	40.33
2006-07	39.12	103.89	34.63
2007-08	24.72	90.53	30.18
2008-09	26.69	111.07	37.02
2009-10	59.66	110	36.67
2010-11	23.65	111.67	37.22
2011-12	28.36	85.32	28.44
2012-13	33.31	93.27	31.09
2013-14	31.60	101.39	33.80
2014-15	36.48	-	-

Moving averages - even number of years (4 years)

□ Add up the values of the first 4 years and place the sum against the middle of 2nd and 3rd year. (This sum is called 4 year moving total)

- ☐ Leave the first year value and add next 4 values from the 2nd year onward and write the sum against its middle position.
- ☐ This process must be continued till the value of the last item is taken into account.
- ☐ Add the first two 4-years moving total and write the sum against 3rd year.
- ☐ Leave the first 4-year moving total and add the next two 4-year moving total and place it against 4th year.
- ☐ This process must be continued till all the 4-yearly moving totals are summed up and centered.
- ☐ Divide the 4-years moving total by 8 to get the moving averages which are our required trend values.

Example 6

Compute the trends by the method of moving averages, assuming that 4-year cycle is present in the following series.

Year	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Annual value	154.0	140.5	147.0	148.5	142.9	142.1	136.6	142.7	145.7	145.1	137.8

Solution:

The four year moving averages are shown in the last column.

Year	Annual value	4-year moving total	Centered total	4-year moving average
1998	154.0			
		-		
1999	140.5		-	
		590.0		
2000	147.0		1168.9	146.11
		578.9		
2001	148.5		1159.4	144.93
		580.5		
2002	142.9		1150.6	143.83
		570.1		
2003	142.1		1134.4	141.8
		564.3		
2004	136.6		1131.4	141.43
		567.1		
2005	142.7		1137.2	142.15
		570.1		
2006	145.7		1141.4	142.68
		571.3		
2007	145.1		-	
		-		
2008	137.8			

4. Method least of squares

Among the four components of the time series, secular trend represents the long term direction of the series. One way of finding the trend values with the help of mathematical technique is the method of least squares. This method is most widely used in practice and in this method the sum of squares of deviations of the actual and computed values is least and hence the line obtained by this method is known as the line of best fit.

It helps for forecasting the future values. It plays an important role in finding the trend values of economic and business time series data.

Computation of Trend using Method of Least squares

Method of least squares is a device for finding the equation which best fits a given set of observations.

Suppose we are given n pairs of observations and it is required to fit a straight line to these data. The general equation of the straight line is:

$$y = a + bx$$

where a and b are constants. Any value of a and b would give a straight line, and once these values are obtained an estimate of y can be obtained by substituting the observed values of y . In order that the equation $y = a + b x$ gives a good representation of the linear relationship between x and y , it is desirable that the estimated values of y_i , say \hat{y}_i on the whole close enough to the observed values $y_i, i = 1, 2, \dots, n$. According to the principle of least squares, the best fitting equation is obtained by minimizing the sum of squares of differences

$$\sum_{i=1}^n \left(y_i - \hat{y}_i \right)^2$$

$$\text{That is, } \sum_{i=1}^n \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^n \left(y_i - a - bx_i \right)^2$$

is minimum. This leads us to two normal equations.

$$\sum_{i=1}^n y_i = na + b \sum_{i=1}^n x_i \quad (7.1)$$

$$\sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 \quad (7.2)$$

Solving these two equations we get the values for a and b and the fit of the trend equation (line of best fit):

$$y = a + bx \quad (7.3)$$

Substituting the observed values x_i in (7.3) we get the trend values $y_i, i = 1, 2, \dots, n$.

Note: The time unit is usually of uniform duration and occurs in consecutive numbers. Thus, when the middle period is taken as the point of origin, it reduces the sum of the time

variable x to zero $\left(\sum_{i=1}^n x_i = 0 \right)$ and hence we get

$$a = \frac{\sum_{i=1}^n y_i}{n} \text{ and } b = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

by simplifying (7.1) and (7.2)

The number of time units may be even or odd, depending upon this, we follow the method of calculating trend values using least square method.

Merits

- ☐ The method of least squares completely eliminates personal bias.
- ☐ Trend values for all the given time periods can be obtained
- ☐ This method enables us to forecast future values.

Demerits

- ☐ The calculations for this method are difficult compared to the other methods.
- ☐ Addition of new observations requires recalculations.
- ☐ It ignores cyclical, seasonal and irregular fluctuations.
- ☐ The trend can be estimated only for immediate future and not for distant future.

Steps for calculating trend values when n is odd:

- (i) Subtract the first year from all the years (x)
- (ii) Take the middle value (A)

(ii) Find $u_i = x_i - A$

(iv) Find u_i^2 and $u_i y_i$

Then use the normal equations:

$$\sum_{i=1}^n y_i = na + b \sum_{i=1}^n u_i$$

$$\sum_{i=1}^n u_i y_i = a \sum_{i=1}^n u_i + b \sum_{i=1}^n u_i^2$$

$$\text{Find } a = \frac{\sum_{i=1}^n y_i}{n} \text{ and } b = \frac{\sum_{i=1}^n u_i y_i}{\sum_{i=1}^n u_i^2}$$

Then the estimated equation of straight line is:

$$y = a + b u = a + b (x - A)$$

Example 7

Fit a straight line trend by the method of least squares for the following consumer price index numbers of the industrial workers.

Year	2010	2011	2012	2013	2014
Index number	166	177	198	221	225

Solution:

Year	Index Number	$X = x_i - 2010$	$u_i = X - A$ $= X - 2$	u_i^2	$u_i y_i$	Trend
2010	166	0	-2	4	-332	165
2011	177	1	-1	1	-177	181.2
2012	198	2	0	0	0	197.4
2013	221	3	1	1	221	213.6
2014	225	4	2	4	450	229.8
	$\sum_{i=1}^5 y_i = 987$		$\sum_{i=1}^5 u_i = 0$	$\sum_{i=1}^5 u_i^2 = 10$	$\sum_{i=1}^5 u_i y_i = 162$	

The equation of the straight line is $y = a + bx$

$= a + bu$ where $u = X - 2$

The normal equations give:

$$a = \frac{\sum_{i=1}^n y_i}{n} = \frac{987}{5} = 197.4$$

$$b = \frac{\sum_{i=1}^n u_i y_i}{\sum_{i=1}^n u_i^2} = \frac{162}{10} = 16.2$$

$$y = 197.4 + 16.2(X - 2)$$

$$= 197.4 + 16.2X - 32.4$$

$$= 16.2X + 165$$

That is, $y = 165 + 16.2X$

To get the required trend values, put $X = 0, 1, 2, 3, 4$ in the estimated equation.

$$X = 0, y = 165 + 0 = 165$$

$$X = 1, y = 165 + 16.2 = 181.2$$

$$X = 2, y = 165 + 32.4 = 197.4$$

$$X = 3, y = 165 + 48.6 = 213.6$$

$$X = 4, y = 165 + 64.8 = 229.8$$

Hence, the trend values for 2010, 2011, 2012, 2013 and 2014 are 165, 181.2, 197.4, 213.6 and 229.8 respectively.

Steps for calculating trend values when n is even:

i). Subtract the first year from all the years (x)

ii). Find $u_i = 2X - (n - 1)$

iii). Find u_i^2 and $u_i y_i$

Then follow the same procedure used in previous method for odd years

Example 8

Tourist arrivals (Foreigners) in Tamil Nadu for 6 consecutive years are given in the following table. Calculate the trend values by using the method of least squares.

Year	2005	2006	2007	2008	2009	2010
No. of arrivals (in lakhs)	12	13	18	20	24	28

Solution:

Year x	No. of arrivals y_i	$X = x_i - 2005$	$u_i = 2X - 5$	u_i^2	$u_i y_i$
2005	12	0	-5	25	-60
2006	13	1	-3	9	-39
2007	18	2	-1	1	-18
2008	20	3	1	1	20
2009	24	4	3	9	72
2010	28	5	5	25	140
	$\sum_{i=1}^6 y_i = 115$		$\sum_{i=1}^6 u_i = 0$	$\sum_{i=1}^6 u_i^2 = 70$	$\sum_{i=1}^6 u_i y_i = 115$

The equation of the straight line is $y = a + bx$

$= a + bu$ where $u = 2X - 5$

Using the normal equation we have,

$$a = \frac{\sum_{i=1}^n y_i}{n} = \frac{115}{6} = 19.17$$

$$b = \frac{\sum_{i=1}^n u_i y_i}{\sum_{i=1}^n u_i^2} = \frac{115}{70} = 1.64$$

$$y = a + bu$$

$$= 19.17 + 1.64 (2X - 5)$$

$$= 19.17 + 3.28X - 8.2$$

$$= 3.28X + 10.97$$

$$\text{That is, } y = 10.97 + 3.28X$$

To get the required trend values, put $X = 0, 1, 2, 3, 4, 5$ in the estimated equation. Thus,

$$X = 0, y = 10.97 + 0 = 10.97$$

$$X = 1, y = 10.97 + 3.28 = 14.25$$

$$X = 2, y = 10.97 + 6.56 = 17.53$$

$$X = 3, y = 10.97 + 9.84 = 20.81$$

$$X = 4, y = 10.97 + 13.12 = 24.09$$

$$X = 5, y = 10.97 + 16.4 = 27.37$$

Hence, the trend values for 2005, 2006, 2007, 2008, 2009 and 2010 are 10.97, 14.25, 17.53, 20.81, 24.09 and 27.37 respectively

Seasonal variation

Seasonal variations are fluctuations within a year over different seasons.

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

Methods of constructing seasonal indices

There are four methods of constructing seasonal indices.

1. Simple averages method
2. Ratio to trend method
3. Percentage moving average method
4. Link relatives method

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

Method Averages Simple

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

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

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

Method of calculating seasonal indices

- (i) The data is arranged season-wise
- (ii) The data for all the 4 seasons are added first for all the years and the seasonal averages for each year is computed.
- (iii) The average of seasonal averages is calculated
(*i.e.*, Grand average = Total of seasonal averages /number of years).
- (iv) The seasonal average for each year is divided by the corresponding grand average and the results are expressed in percentages and these are called seasonal indices.

Example 9

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

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

Solution:

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

$$\begin{aligned}
 \text{Grand Average} &= \frac{\text{Total of seasonal averages}}{4} \\
 &= \frac{169.82 + 290.22 + 498.1 + 24.18}{4} \\
 &= \frac{982.32}{4} = 245.58
 \end{aligned}$$

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

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

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

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

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

Example 10

Obtain the seasonal indices for the rain fall (in mm) data in India given in the following table.

Quarter \ Year	2009	2010	2011	2012
I	38.2	38.5	55	50.5
II	166.8	250.9	277.7	197
III	612.6	773.1	717.8	706.1
IV	72.2	153.1	65.8	101.1

Solution:

Year	Quarter			
	I	II	III	IV
2009	38.2	166.8	612.6	72.2
2010	38.5	250.9	773.1	153.1
2011	55	277.7	717.8	65.8
2012	50.5	197	706.1	101.1
Seasonal total	182.2	892.4	2809.6	392.2
Seasonal average	45.55	223.1	702.4	98.05
Seasonal index	17	83	263	37

$$\begin{aligned}
 \text{Grand Average} &= \frac{\text{Total of seasonal averages}}{4} \\
 &= \frac{45.55 + 223.1 + 702.4 + 98.05}{4} \\
 &= \frac{1069.10}{4} = 267.28
 \end{aligned}$$

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

$$\text{Seasonal Index for Quarter I} = \frac{45.55}{267.28} \times 100 = 17.04 \approx 17$$

$$\text{Seasonal Index for Quarter II} = \frac{223.1}{267.28} \times 100 = 83.47 \approx 83$$

$$\text{Seasonal Index for Quarter III} = \frac{702.4}{267.28} \times 100 = 262.80 \approx 263$$

$$\text{Seasonal Index for Quarter IV} = \frac{98.05}{267.28} \times 100 = 36.69 \approx 37$$

QUESTION BANK
UNIT V TIME SERIES

PART A

1. Define time series
2. What are the components of time series.
3. Lists the method of constructing seasonal indices.
4. Draw diagram for cyclic variation.
5. State secular trend in time series.
6. State Additive approach in time series
7. State Multiplicative approach in time series
8. Lists difference between additive and multiplicative approach
9. What are methods of measuring trends in time series.
10. State Method of least squares.

PART B

1. Explain different components of time series.
2. Compose any two methods of measuring trends
3. Elaborate different methods of constructing seasonal indices.
4. Calculate the trend values using semi-averages methods for the income from the forest department. Find the yearly increase.

Year	2008	2009	2010	2011	2012	2013
Income (in crores)	46.17	51.65	63.81	70.99	84.91	91.64

5. Calculate the 3-year moving averages for the loans issued by co-operative

banks for non-farm sector/small scale industries based on the values given below.

Year	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15
Amount	41.82	40.05	39.12	24.72	26.69	59.66	23.65	28.36	33.31	31.60	36.48

6. Construct various seasonal indices methods in time series.
7. Compare simple average method and trend methods.
8. Tourist arrivals (Foreigners) in Tamil Nadu for 6 consecutive years are given in the following table. Calculate the trend values by using the method of least squares.

Year	2005	2006	2007	2008	2009	2010
Arrivals	12	13	18	20	24	28

9. Explain computation of trend using least square methods.
10. Illustrate Moving Average Methods in time series.

