UNIT 3 STATISTICS

Structure

3.1 Introduction

Objectives

3.2 Statistical Data and Variables and Units of Observations

3.2.1 Raw Data

3.2.2 Variables of Observation

3.2.3 Unit of Observation

3.3 Construction of Frequency Tables (or Frequency Distributions) from Raw Data

3.4 Graphical Presentation of Frequency Distributions

3.5 Measures of Location and Dispersion

3.5.1 Measures of Location

3.5.2 Measures of Dispersion

3.6 Development of Probability Theory

3.6.1 Basic Concepts of Probability

3.6.2 Some Useful Terms

3.7 Theorems of Probability

3.8 The Principle of Additivity

3.9 Conditional Probability

3.10 Bayes's Theorem

3.11 Random Variables

3.12 Distribution Functions

3.12.1 The Binomial Distribution

3.12.2 The Poisson Distribution

3.12.3 The Normal Distribution

3.13 Statistical Inference and Sample Statistics

3.14 Correlation

3.15 Regression

3.16 Summary

3.17 Key Words

3.18 Answers to SAQs

3.1 INTRODUCTION

The word 'statistics' appears to have been derived from the Latin word 'status' meaning a (political) state. In its origin, statistics was simply the study of "The political arrangement of the modern states of the known world". The description of states was at first verbal but later the increasing proportion of numerical data in the description gradually gave the word 'statistics'. The scope of statistics now includes collection of numerical data pertaining to almost every field; for this reason it is very useful in economics, sociology, business, education, agriculture, psychology, biology and related fields. It may be defined as a science which enables us to draw representative samples, analyse the data collected, interpret and make inferences.

`It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge."

Objectives

After studying this unit, you should be able to

- construct the frequency tables given a numerical data,
- define the measures of location namely mean and the median and analyse the information it conveys,
- define the measures of dispersion namely the standard deviation, and the mean deviation and analyse the information it conveys,
- evaluate the probability of certain combination of events and combination probability,
- apply Bayes's theorem to find the probability,
- identify a discrete random variable,
- obtain probability distribution in some simple cases, and
- evaluate mean, variance and standard deviation for some simple probability distribution device.

3.2 STATISTICAL DATA AND VARIABLES AND UNITS OF OBSERVATIONS

3.2.1 Raw Data

Statistics is a collection of information in numerical terms. For example, marks obtained by the students of a class; monthly wages of workers in a factory; numbers indicating births, deaths and marriages in different states etc. are called Statistical Data or Numerical Data and statistics is the science which deals with the collection, analysis and interpretation of statistical data.

The numerical data or information is collected in two ways. When information is collected in respect of every individual person or item, then the numerical data has been carried out by means of complete Enumeration or Census. But if information is collected only from a selected portion (or sample) of a given population, the procedure is called Sample Survey.

For example, during the census operation, the population of a country is enumerated and all citizens are included in such operation but while verifying the accuracy of entries in Books of Accounts, we check only a portion of entries. This is an example of Sample Survey.

In addition to these two methods, we also come across regular collection and recording of information in a routine manner for example the Railways keep a daily record of movement of passengers and goods, the income earned through fares and freights etc.

The information collected through censuses and surveys or in a routine manner is called raw data.

3.2.2 Variables of Observation

In a census, suppose for each person we have recorded the age and sex and whether the person belongs to a rural or urban area, we say that we have taken observations on three variables : age, sex and place of residence. The term variable stands for what is being observed. A variable is completely described by its descriptive name and the description

of all the values it can possibly take. In the above example, the variable called 'place of residence' has two possible values as 'rural' or 'urban'. The value of this variable could also have been recorded as the name of the state, district, city, village etc. Thus, the variable 'place of residence' has the same name but the set of values may be different for them. We regard two variables to be different if their set of possible values are different even if they have the same name.

Qualitative and Quantitative Variables

Variables of observations with numbers as possible values are called quantitative variables and variables of observation with name of things, places etc. are called qualitative variables. For example in the example of census, the value of age variable were numbers so the age variable is a quantitative variable whereas the variable called 'place of residence' is a qualitative variable. A word of caution is necessary here. Suppose in the recording of the variable 'place of residence' we mark '1' if urban and '2' if rural but this does not make the variable a quantitative variable. A quantitative variable has not only its values recorded as numbers but they are really number on which arithmetic operations can be carried out. For example, age is a quantitative variable as sum, product, difference of age has a sense whereas the sum of rural or urban 'place of residence' has no sense.

Age, height, income of a worker etc. are examples of quantitative variables whereas variables such as sex, religion, caste etc. are qualitative variables.

3.2.3 Unit of Observation

Suppose in the census example we have recorded age, sex and place of residence of all the persons alive at the time of the census and in another example we have recorded the results of a particular examination, i.e. we have recorded the name of the students who appeared in that examination and marks obtained by each student.

The term unit of observation will be used to describe what the values of a variable are attached to. In the census example, the unit of observations are person alive at the time of census and to each unit of observation we recorded the value of three variables : age, sex and place of residence. In the example of results of a particular examination, the unit of observations would be the students who have appeared in the said examination and the variable of observation will be marks obtained by a particular student. Thus different variables of observation may be associated with the same unit of observation.

3.3 CONSTRUCTION OF FREQUENCY TABLES (OR FREQUENCY DISTRIBUTIONS) FROM RAW DATA

If a sample of a population contains a 'large' number of observations, the investigator has to devise methods to condense it and present it in the form of tables and charts to bring out its main characteristics. This is called **data presentation**.

Let us consider, for example, the marks obtained by 30 students of Class XI in a class test (out of 80 marks) in mathematics.

60, 49, 53, 57, 73, 62, 40, 39, 68, 55, 36, 61, 40, 31, 43, 41, 47, 52, 67, 44, 54, 52, 24, 38, 48, 46, 72, 46, 47, 49.

These observations constitute **raw data** or **ungrouped data**. What do these 30 numbers convey to us? Not much. We, therefore, would like to bring out certain features of this data. For instance, we could arrange numbers in ascending or descending order of magnitude. But, this method would involve difficulties when the number of observations is very large. So, what we generally do is, to condense the data into **classes** (or **groups**) as follows :

We find the difference between maximum and minimum observations. This difference is called **range** of the raw data. Then, we decide about the number of classes into which the raw data is to be grouped. Care should be taken that the classes cover the entire range and there is a class to include the least observation and also a class to include greatest

observation. In general, we make sure that we have **not less than 5 or more than 15 classes**.

In the above example, the range is 73 – 24 = 49. So, it is convenient to have 10 classes, each of width (or size) 5. In general, the width of each class (or class-interval) is a convenient whole number immediately greater than the quotient obtained by dividing the range by the number of classes to be made. While setting up the class limits (i.e. the maximum and minimum numbers which can be put in the class), the following rules may be observed :

(i) Classes should be non-overlapping.

(ii) The classes should be continuous without any gap.

(iii) As far as possible, the classes should be of the same size.

(iv) Classes such as, less than 3 or more than 8, i.e. open-ended classes should be avoided.

(v) The boundaries of each class should be determined in such a way that there is no ambiguity as to which class a particular observation of the data belongs.

In our example, we choose the classes as

24 – 29, 29 – 34, 34 – 39, . . . , 69 – 74.

Note that the class 24 – 29 will contain all observations which are greater than or equal to 24 but less than 29. The observation '29' will be put in the next class and so on.

24 is the **lower limit** of the class 24 – 29 and 29 is the **upper limit** of the class. The arithmetic average of the lower limit and the upper limit of a class is called the **class-mark** of that class. The class mark of the class 24 – 29 is $26.5\left(=\frac{24+29}{2}\right)$.

To prepare the frequency distribution table, we take each observation from the data, one at a time and put a tally mark (a right-handed dash) opposite the class in which the observation lies. For sake of convenience and symmetry, we record tally marks in bunches of five, the fifth one crossing the other four diagonally. The count of tally marks in a particular class is called **frequency** of that class and is recorded opposite the class next to the tally marks. It may be noted that the sum of all the frequencies is equal to the total number of observations in the raw data.

Table 3.1

Class-interval (Marks out of 80)	Tally Marks	Frequency
24 – 29	/	1
29 – 34	/	1
34 – 39	//	2
39 – 44	////	5
44 – 49	//// /	6
49 – 54	////	5
54 – 59	///	3
59 – 64	///	3
64 – 69	//	2
69 – 74	//	2
Total		30

Using the above steps, the **frequency distribution table** of the marks obtained by 30 students of class XI in a class test (out of a maximum of 80) in mathematics is as shown in Table 3.1.

The data in the above form are called **grouped data**. We have condensed 30 observations into ten classes and we observe from this data that

(i) There are 4 students (out of 30) who have secured less than 39 (of course, ≥ 24) marks.

(ii) Nearly half the students (16) have secured marks between 39 and 54.

(iii) Only two students have secured 69 or more than 69 (of course, ≤ 74) marks.

Often, we shall be interested in knowing the number of observations less than a particular number. For this purpose we add another column in the above frequency table. Opposite to each class we write in this column, the sum of frequencies of all the previous classes and that particular class. The new number we get is called the **cumulative frequency** of the class and the modified table is called **cumulative frequency table**.

Surely, the cumulative frequency of the last class is same as the total number of observations in the data. The cumulative frequency table in case of our example is the Table 3.2.

Table 3.2

Class-interval (Marks out of 80)	**Tally Marks**	**Frequency**	**Cumulative Frequency**
24 – 29	/	1	1
29 – 34	/	1	2
34 – 39	//	2	4
39 – 44	~~////~~	5	9
44 – 49	//// /	6	15
49 – 54	////	5	20
54 – 59	///	3	23
59 – 64	///	3	26
64 – 69	//	2	28
69 - 74	//	2	30
Total		30	

From this table we can say, at a glance, that 15 (our of 30) students have secured less than 49 marks and only 4 students have secured less than 39 marks.

In fact, there are two types of cumulative frequencies *upward* or *downwards*. A cumulative frequency distribution may be made on a '*less than*' basis or on a '*more than*' basis. Table 3.2 has been constructed on a 'less than' basis. For comparison, we give below, both the tables, one constructed on less than basis and the other on more than basis.

Table 3.3 (i)
Cumulative Frequency Table Constructed on Less than Basis

Score	Number of Students Scoring Less than the Indicated Score
24	0
29	1
34	2
39	4
44	9
49	15
54	20
59	23
64	26
69	28
Total	30

Table 3.3 (ii)
Cumulative Frequency Table Constructed on More than Basis

Score	Number of Students Scoring More than or Equal to the Indicated Score
24	30
29	29
34	28
39	26
44	21
49	15
54	10
59	7
64	4
69	2
Total	0

In practical situations, it is often desired to compare class frequencies in two or more distributions based upon a very different number of total items. This becomes very easy, if we transform the absolute frequencies into relative frequencies. The class frequencies expressed *relative* to the total frequency (total number of items) are called **percentage frequencies**. We arrive at percentage frequencies by dividing the frequencies in each class by the total number of items in the distribution and express the resulting fraction as percents. The following table illustrates the whole process.

Table 3.4

Class	Frequency	Cumulative Frequency
24 – 29	1	$\frac{1}{30} = 0.0333 = 3.33\%$
29 – 34	1	$\frac{1}{30} = 0.0333 = 3.33\%$
34 – 39	2	$\frac{2}{30} = 0.0667 = 6.67\%$
39 – 44	5	$\frac{5}{30} = 0.1667 = 16.67\%$
44 – 49	6	$\frac{6}{30} = 0.2000 = 20.00\%$
49 – 54	5	$\frac{5}{30} = 0.1667 = 16.67\%$
54 – 59	3	$\frac{3}{30} = 0.1000 = 10.00\%$
59 – 64	3	$\frac{3}{30} = 0.1000 = 10.00\%$
64 – 69	2	$\frac{2}{30} = 0.0667 = 6.67\%$
69 – 74	2	$\frac{2}{30} = 0.0667 = 6.67\%$
Total	30	

Sometimes, the classes are not continuous, i.e. the upper class limit of a class is not equal to the lower class-limit of the next class; then we make the classes continuous by decreasing lower limit of each class by 0.5 and increasing upper limit by 0.5. Consider the example of distribution of ages (in years) of primary school teachers in a Tehsil :

Table 3.5(a)

Age (in Years)	Number of Primary Teachers
21 – 25	20
26 – 30	26
31 – 35	34
36 – 40	47
41 – 45	15
46 – 50	7
51 – 55	3
Total	152

We shall modify the above frequency distribution table as follows :

Table 3.5(b)

Age (in Years)	Number of Primary Teachers
20.5 – 25.5	20
25.5 – 30.5	26
30.5 – 35.5	34
35.5 – 40.5	47
40.5 – 45.5	15
45.5 – 50.5	7
50.5 – 55.5	3
Total	152

Example 3.1

Given the following distribution of weekly wage rates of a selected group of junior and senior typists in a private factory; compare the two distributions by constructing relative frequency distributions :

Weekly Wage Rate (in Rupees)	Number of Junior Typists	Number of Senior Typists
100-200	12	
200-300	32	
300-400	38	16
400-500	30	19
500-600	13	10
600-700		3
700-800		2

Solution

We construct the table of relative frequencies as follows :

Table 3.6

Weekly Wage Rate (in Rupees)	Number of Typists		Percent of Total Number of Typists	
	Junior	Senior	Junior	Senior
100-200	12		$\frac{1200}{125} = 9.6$	
200-300	32		$\frac{3200}{125} = 25.6$	
300-400	38	16	$\frac{3800}{125} = 30.4$	$\frac{1600}{50} = 32.0$
400-500	30	19	$\frac{3000}{125} = 24.0$	$\frac{1900}{50} = 38.0$
500-600	13	10	$\frac{1300}{125} = 10.4$	$\frac{1000}{50} = 20.0$
600-700		3		$\frac{300}{50} = 6.0$
700-800		2		$\frac{200}{50} = 4.0$
Total	125	50	100	100

3.4 GRAPHICAL PRESENTATION OF FREQUENCY DISTRIBUTIONS

The main features of frequency distribution are conveniently communicated by representing the frequency distribution in the term of a diagram, since a diagram is more easily and more quickly understood than a collection of numbers. Diagrammatic presentation is particularly useful when the number of classes or the class frequency distribution is large.

There are various methods of graphical presentation of frequency distribution which are in use. We shall discuss only two of them namely the bar diagram and the pie diagram.

The Bar Diagram

To draw the bar diagram of a frequency distribution, we mark **equal lengths** on horizontal axis for representing difference classes. These lengths must be equal even if the classes are of unequal size. On each of these lengths (on the horizontal axis), we erect a *rectangle* whose height is proportional to the frequency of the class represented by its base. This means that the heights of a rectangle represents the relative frequency of the class represented by its base. Thus, we shall get '*bars*' and hence the name bar diagram. The bar diagram of frequency distribution of Table 3.1 (using the Table 3.4) is shown in Figure 3.1.

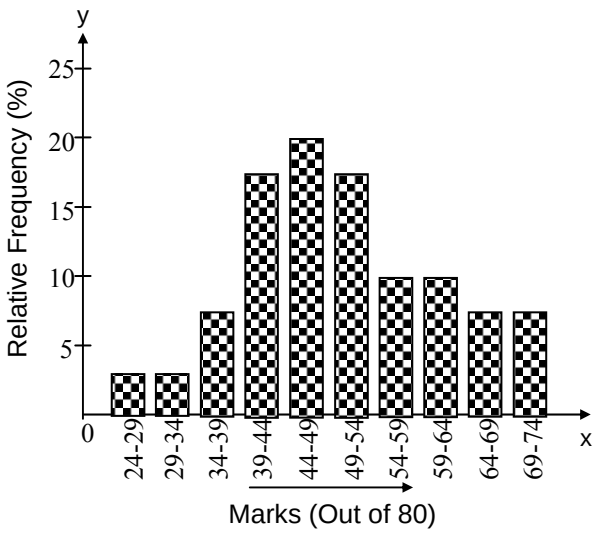


Figure 3.1

The bar diagram (or bar chart) is particularly useful when it is desired to compare two different frequency distributions. This is achieved by drawing two bars (for the same class) adjacent to each other, one for the first distribution and the other for the second, their heights representing the relative frequencies in their respective distribution. The bar diagram for studying the comparison of two distribution of Table 3.6 (Example 3.1) is shown in Figure 3.2.

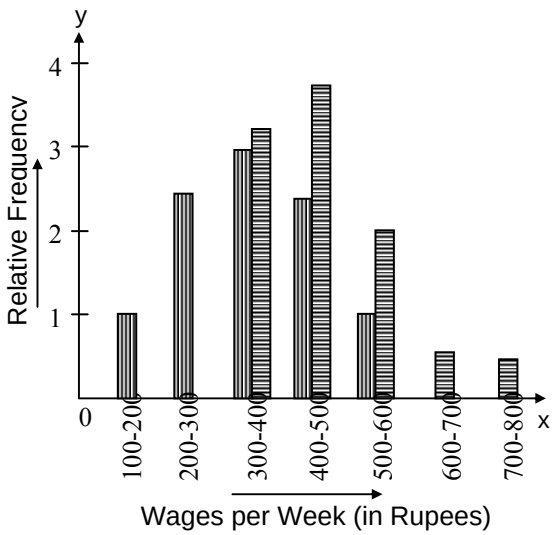


Figure 3.2

The Pie Diagram

The pie diagram (or pie chart) is used to represent relative frequencies only. The relative frequencies of different classes are represented by sectors of a circle. The angle of each sector is proportional to the relative frequency of the particular class represented by the sector. The angle of a sector representing a particular class is calculated by multiplying 360° with the relative frequency of that class. The angles of sectors representing the frequency distribution of Table 3.1 is calculated in the following table using Table 3.4.

The pie diagram of the distribution (Table 3.7) is shown in Figure 3.3.

Table 3.7

Class	Percentage Frequency	Angle of the Sector Representing the Class
24 – 29	3.33	$\frac{3.33 \times 360^o}{100} = 12^o$
29 – 34	3.33	$\frac{3.33 \times 360^o}{100} = 12^o$
34 – 39	6.67	$\frac{6.67 \times 360^o}{100} = 24^o$
39 – 44	16.67	$\frac{16.67 \times 360^o}{100} = 60^o$
44 – 49	20.00	$\frac{20 \times 360^o}{100} = 72^o$
49 – 54	16.67	$\frac{16.67 \times 360^o}{100} = 60^o$
54 – 59	10.00	$\frac{10 \times 360^o}{100} = 36^o$
59 – 64	10.00	$\frac{10 \times 360^o}{100} = 36^o$
64 – 69	6.67	$\frac{6.67 \times 360^o}{100} = 24^o$
69 – 74	6.67	$\frac{6.67 \times 360^o}{100} = 24^o$
Total	100	360°

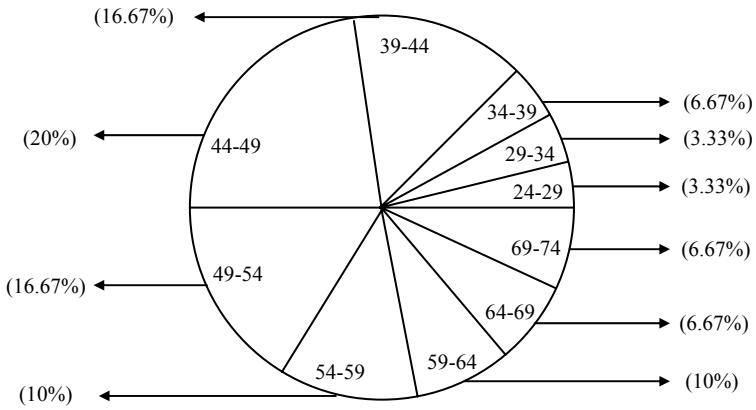


Figure 3.3 : Pie Diagram Showing the Frequency Distribution of Table 3.7

3.5 MEASURES OF LOCATION AND DISPERSION

So far we have discussed the presentation of raw data in a form suitable for communicating the information contained in it and have studied the use of frequency tables. In case of quantitative variables the information contained in the raw data can be summarized by means of a few numerical values. Such a summary is partly provided by what are called measures of location and measures of dispersion.

3.5.1 Measures of Location

Definition 1

The Arithmetic Mean of the values $x_1, x_2, \ldots, x_n$ of a variable recorded for n units of observation is defined as $\dfrac{x_1 + x_2 + \ldots + x_n}{n}$ *and is denoted by* $\bar{x}$.

$$\therefore \qquad \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

From the definition of $\bar{x}$, we have

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

i.e. $\quad (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0$

Thus some values of $x_i - \bar{x}$ must be positive and some negative so that the sum is zero. If we add all the positive $x_i - \bar{x}$ and all the negative $x_i - \bar{x}$, these two sums will have the same value but opposite in sign so that their algebraic sum is zero.

Hence we say that $\bar{x}$ lies at the centre of all the observations. Unless otherwise stated, the word mean denotes the arithmetic mean.

For calculating the mean of the group data, suppose the observed values in the different classes of the frequency table are $y_1, y_2, \ldots, y_k$ and the frequencies in each class are $f_1, f_2, \ldots, f_k$, then mean will be given by

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i\, y_i}{\sum_{i=1}^{k} f_i}$$

In some cases the classes are no longer defined by single values, each class consists of many values of the variable. In such a case the method is to replace all the observed values belonging to a class by the mid value of that class and then use the mid value to determine the mean. In this method one of the class marks (preferably near the middle) is designated as 'a' (called the assumed mean) and the deviation $d_i = y_i - a$ are calculated for each class.

The arithmetic mean is then

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i\, y_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i\,(d_i + a)}{\sum_{i=1}^{k} f_i}$$

$$\bar{x} = \frac{\sum_{i=1}^{k} (f_i d_i + f_i a)}{\sum_{i=1}^{k} f_i} = a + \frac{1}{n} \sum_{i=1}^{k} f_i d_i$$

where k is the number of classes and n, the number of observations,

i.e. $$n = \sum_{i=1}^{k} f_i$$

i.e. $$\bar{x} = a + \frac{1}{n} \sum_{i=1}^{k} f_i \, d_i$$

Since in most of the problems, the width of all the classes is same, we can further simplify the calculations of the mean of the grouped data by calculating the mean of μ_i's (denoted by Mu), where $\mu_i = \frac{y_i - a}{c}$, c is the width of each class and a is an assumed mean.

Now, $$\bar{x} = a + \frac{1}{n} \sum_{i=1}^{k} f_i \, d_i$$

$$= a + \frac{1}{n} \sum_{i=1}^{k} c f_i \, . \frac{d_i}{c}$$

$$= a + \frac{c}{n} \sum_{i=1}^{k} f_i \, . \frac{d_i}{c}$$

$$= a + c \left[\frac{1}{n} \sum_{i=1}^{k} f_i \, \mu_i \right]$$

$$= \left(\Theta \; \frac{d_i}{c} = \frac{y_i - a}{c} = \mu_i \right)$$

$$= a + c \, \mu$$

This method is known as the step deviation method.

Remark

The step deviation method works equally well if the classes are of unequal width. The only care to be taken is that the number c (which may not be the class size) should be a divisor of each class size.

Example 3.2

The marks obtained by 20 students in a test were 13, 17, 11, 5, 18, 16, 11, 14, 13, 12, 18, 11, 9, 6, 8, 17, 21, 22, 7, 6.

Find (i) The mean marks per student.

(ii) The mean marks per student when marks of each student are increased by 5.

(iii) The mean marks per student when the marks of each student are doubled.

Solution

The sum of marks of all the students = 13 + 17 + 11 + 5 + 18 + 16 + 11 + 14 + 13 + 12 + 18 + 11 + 9 + 6 + 8 + 17 + 21 + 22 + 7 + 6 = 255.

(i) $\text{Mean} = \dfrac{\sum x_i}{\text{Number of students}} = \dfrac{255}{20} = 12.75$

(ii) When marks of each student is increased by 5, then sum of their marks is increased by 20 × 5 = 100, i.e. sum of marks = 255 + 100 = 355.

$$\text{Mean} = \frac{355}{20} = 17.75$$

Thus we see that mean is also increased by 5.

(iii) When marks of each student is doubled, the sum of their marks is also doubled, i.e. the sum of the marks = 255 × 2 = 510.

$$\text{Mean} = \frac{510}{20} = 25.5 = 2 \times 12.75$$

i.e. the mean has also doubled.

Example 3.3

The following table shows the gain in weight by 25 children in a year

Gain in Weight (in kg)	2	2.5	3	3.5	4	4.5	5	5.5	6
No. of Children	2	3	4	2	5	1	4	3	1

Find the mean of gain in weight.

Solution

For calculation of the mean, we construct the table

y_i	f_i	$f_i y_i$
2.0	2	4.0
2.5	3	7.5
3.0	4	12.0
3.5	2	7.0
4.0	5	20.0
4.5	1	4.5
5.0	4	20.0
5.5	3	16.5
6.0	1	6.0
Total	25	97.5

$$\text{Mean} = \frac{\sum f_i\, y_i}{\sum f_i} = \frac{97.5}{25} = 3.9$$

So mean of gain in weight = 3.9.

Example 3.4

The weekly observations on cost of living index in a certain city for a particular year are

Cost of Living Index	140-150	150-160	160-170	170-180	180-190	190-200

Number of Weeks	5	10	20	9	6	2

Compute the average weekly cost of Living Index.

Solution

We shall use the deviation method by taking a = 165 the assumed mean.

Class	Class Mark (y_i)	Frequency (f_i)	$d_i = y_i - a$	$f_i\, d_i$
140 – 150	145	5	– 20	– 100
150 – 160	155	10	– 10	– 100
160 – 170	165	20	0	+ 0
170 – 180	175	9	10	+ 90
180 – 190	185	6	20	120
190 – 200	195	2	30	60
Total		52		70

$$\therefore \quad \text{Mean } = a + \frac{\sum f_i\, d_i}{\sum f_i} = 165 + \frac{70}{52}$$

$$= 165 + 1.35 = 166.35$$

Example 3.5

The ages of all the male inhabitants of a village were received and the following frequencies distribution was obtained.

Age (years)	0-5	5-10	10-20	20-30	30-40	40-50	50-60	60-80
Number of Persons	12	18	16	19	14	11	4	3

Obtain the mean age per male inhabitant.

Solution

We construct the following table, taking assuming mean a = 25, c = 2.5.

Class	Class Mark (y_i)	Frequency ($f_i i$)	$\mu_i = \frac{y_i - a}{c}$	$f_i\, \mu_i$
0 – 5	2.5	12	– 9	– 108
5 – 10	7.5	18	– 7	– 126
10 – 20	15	16	– 4	– 64
20 – 30	25	19	0	0
30 – 40	35	14	4	56
40 – 50	45	11	8	88
40 – 60	55	4	12	48
60 – 80	70	3	18	54
Total		97		– 52

$$\therefore \quad \text{Mean} = a + c \times \frac{\sum f_i\, \mu_i}{\sum f_i}$$

$$\therefore \quad \text{Mean } = 25 + 2.5 \times \frac{-52}{97}$$

$$= 25 - \frac{130}{97} = 25 - 1.34$$

$$= 23.66 \text{ nearly.}$$

Definition 2 : Median

Median is defined as the central value of a set of observations. It divides the whole series of observations into two parts. If there are n observations $x_1, x_2, x_3, \ldots, x_n$, *then*

$$Median = \begin{cases} \frac{n+1}{2} th \text{ observation}, & \text{if } n \text{ is odd} \\ \frac{\frac{n}{2} th \text{ observation} + \left(\frac{n}{2}+1\right) th \text{ observation}}{2}, & \text{if } n \text{ is even} \end{cases}$$

where $x_1, x_2, x_3, \ldots, x_n$ *are either in ascending or in descending order.*

For example, the median of 1, 2, 4, 8, 9, 10, 12 is $\frac{7+1}{2}$ *th* term

$= 4$ *th* term $= 8$

and the median of 3, 5, 8, 9, 12, 15, 16, 18, 19, 23

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

$$= \frac{1}{2}(5th \text{ term} + 6th \text{ term})$$

$$= \frac{1}{2}(12+15) = 13.5.$$

Alternatively, the median of a set of n observations is a number M which satisfies the conditions :

(i) *(Number of observations* $\geq M$*)* $\geq \frac{n}{2}$.

(ii) *(Number of observations* $\leq M$*)* $\geq \frac{n}{2}$.

Consider the set 8, 9, 5, 3, 12, 18, 15, 16, 23, 19 of 10 numbers. The numbers arranged in ascending order are 3, 5, 8, 9, 12, 15, 16, 18, 19, 23.

If M is the median, then there should be atleast $5\left(=\frac{10}{2}\right)$ numbers greater than or equal to M; this suggests that $M \leq 15$. Also, there should be atleast 5 numbers less than or equal to M; this suggests that $M \geq 12$. This means that any number between 12 and 15 can be taken as the median. Conventionally, we take the mean of 12 and 15 as the median.

We observe from the above definition of median that unlike the arithmetic mean, median of a set of observations may not be unique. However, the methods given here are conventional and these determine the median without ambiguity.

In case of *grouped data*, the median is calculated by formula :

$$\text{Median} = l + \frac{\frac{n}{2} - C_{f-1}}{f_m} \times c_m,$$

where l = lower limit of the median class,

f_m = frequency of the median class,

C_{f-1} = cumulative frequency of the class preceding to the median class,

c_m = width of the median class, and

n = sum of all the frequencies, i.e. total number of observations.

The **median class** being the class which contains the $\frac{n}{2}th$ observation.

Remark

The above method of finding the median in case of grouped data works well even if the classes are of unequal widths. Of course, we assume that the **frequency of the median class is uniformly distributed over the whole class** and the classes are without gaps and they have been arranged according to the ascending order of the variable.

Example 3.6

The number of students absent in a school was recorded every day for 147 days and the new data was presented in the form of the frequency table given below :

Number of Students Absent	5	6	7	8	9	10	11	12	13	15	18	20
Number of Days	1	5	11	14	16	13	10	70	4	1	1	1

Obtain the median and describe the information conveyed by it.

Solution

Here, n = 147, and odd number, therefore the median $= \frac{n+1}{2}th$, i.e. $\frac{147+1}{2}th$ observations, i.e. the median is 74 *th* observation. To find it, we construct the cumulative frequency table.

x_i	5	6	7	8	9	10	11	12	13	15	18	20
y_i	1	5	11	14	16	13	10	70	4	1	1	1
Cunulative Frequency	1	6	17	31	47	60	70	140	144	145	146	147

We notice that 74 *th* observation is 12.

(Θ all observations from 71st upto 140 are equal, each being 12.)

This value of the median suggests that for half the number of days 12 or more than 12 students remained absent and on the other days 12 or less than 12 students remained absent.

Example 3.7

Calculate the mean and median of the following data :

Number of Workers	12	30	65	107	157	202	222	230
Wages per Week up to (Rs.)	15	30	45	60	75	90	105	120

Solution

In this case, we are given the cumulative frequencies. We construct the following table for mean and the median. Here, width of each class = 15. We shall use step deviation method taking a = 67.5.

Class	Frequency y_i	Cumulative Frequency	f_i	$\mu_i = \frac{y_i - a}{15}$	$f_i\,\mu_i$
0 – 15	7.5	12	12	– 4	– 48
15 – 30	22.5	30	18	– 3	– 54
30 – 45	37.5	65	35	– 2	– 70
45 – 60	52.5	107	42	– 1	– 42
60 – 74	67.5	157	50	0	0
75 – 90	82.5	202	45	1	45
90 – 105	97.5	222	20	2	40
105 – 120	112.5	230	8	3	24
Total			230		– 105

$\therefore$ The mean $= a + c \times \left(\frac{\Sigma f_i \mu_i}{\Sigma f_i}\right) = 67.5 + 15 \times \frac{-105}{230}$

$= 67.5 - 6.85 = 60.65$ nearly.

Here $\frac{n}{2} = \frac{230}{2} = 115$

$\therefore$ Median class is 60 – 75.

$\therefore$ Median $= l + \frac{\frac{n}{2} - C_{f-1}}{f_m} \times c_m$

$= 60 + \frac{115 - 107}{50} \times 15$

$= 60 + 2.4 = 62.4$

3.5.2 Measures of Dispersion

Suppose a cricket team to represent India was to be selected and all the members of the team have been selected except one. Two players *X* and *Y* are available and the last member to be selected has to be one of them. The managers look at the runs scored by these two batsmans in the last 5 matches which are as follows :

X	38, 70	48, 34	42, 55	63, 46	54, 44
Y	5, 11	8, 29	83, 104	20, 28	81, 123

The average score per inning is nearly 50 for both the players. We observe that the runs made by player *X* do not change much from inning to inning whereas *Y*'s scores show great variation with very high scores in one inning and very low in another. We use the word dispersion and say that the runs scored by *Y* show a higher dispersion than the runs made by *X*. The mean of runs made by *X* is 49.4 and his scores are close to the mean score whereas the mean score of *Y* is 49.2 and his scores in different innings are not close to the mean score.

The measure of location, the mean and the median give us a central value around which the values of the variables are located but gives us no idea of how far these values are from the central value. The measure of dispersion which we are going to study now provides us this information.

The commonly used measure of dispersion are Standard Deviation (SD) and the Mean Deviation (MD). Standard deviation is the measure of dispersion about the mean and mean deviation is the measure of dispersion about the median.

Definition 3 : Mean Deviation

The mean deviation is the mean of the absolute differences of the values from the mean or median. Thus mean deviation (*MD*)

$$= \frac{1}{n} \sum f_i \, |x_i - A|$$

where A is either the mean or median. As the positive and negative differences leave equal effects, only the absolute value of differences is taken into account.

Now let us consider the mean deviation from the median.

Mean Deviation from the Median

Mean deviation from the median

$$\text{MD} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{\sum_{i=1}^{n} |d_i|}{n},$$

n being the total number of observations and $\bar{x}$ being the median. In case of a grouped data, the mean deviation about median

$$\text{MD} = \frac{\sum_{i=1}^{k} f_i \, |y_i - \bar{y}|}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i \, |d_i|}{\sum_{i=1}^{k} f_i}$$

where $\quad d_i = y_i - \bar{y}$.

Definition 4 : Standard Deviation

Let us consider a set of n observations $x_1, x_2, x_3, \ldots, x_n$. *We compute the sum S of squares of deviations of these observations from an arbitrary number a.*

So, $\quad S = (x_1 - a)^2 + (x_2 - a)^2 + (x_3 - a)^2 + \ldots + (x_n - a)^2$

$$= \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} [(x_i - \bar{x}) + (\bar{x} - a)]^2$$

where $\bar{x}$ *is the mean of n observations in reference*

$$= \sum_{i=1}^{n} [(x_i - \bar{x})^2 + (\bar{x} - a)^2 + 2(x_i - \bar{x})(\bar{x} - a)]$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - a)^2 + 2(\bar{x} - a) \sum_{i=1}^{n} (x_i - \bar{x})$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - a)^2$$

$$\left[\Theta \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0\right]$$

Clearly, S is minimum when $\bar{x} - a = 0$, i.e. when $\bar{x} = a$, i.e. when the deviations are considered from the arithmetic mean.

In view of the above idea, *Karl Pearson introduced the concept of standard deviation*. It is most popular measure of dispersion. It is denoted by σ and is defined as

(i) $$\sigma = \left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \right]^{\frac{1}{2}} \quad \ldots \text{(i)}$$

in case of ungrouped data, and

(ii) $$\sigma = \left[\frac{\sum_{i=1}^{k} f_i \, (y_i - \bar{y})^2}{\sum_{i=1}^{k} f_i} \right]^{\frac{1}{2}}$$

in case of grouped data, it being assumed that frequency of a class is centred at the class mark.

Calculation for standard deviation can be more simplified if we take deviations of the variates (or class marks) from an assumed mean 'a'.

Let $d_i = x_i - a$, then $x_i = d_i + a$ . . . (ii)

From Eq. (i),

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i^2 + \bar{x}^2 - 2 x_i \bar{x})$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i^2 + \frac{1}{n} (n \bar{x}^2) - \frac{2\bar{x}}{n} \sum_{i=1}^{n} x_i$$

$$\left(\Theta \sum_{i=1}^{n} \bar{x} = \bar{x} + \bar{x} + \bar{x} + \ldots \text{upto } n \text{ terms} = n\bar{x} \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i^2 + \bar{x}^2 - 2 \frac{\bar{x}}{n} (n \bar{x}) \left(\Theta \; \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

or $$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \quad \ldots \text{(iii)}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (d_i + a)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} (d_i + a) \right)^2 \qquad \text{(using (ii))}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (d_i^2 + a^2 + 2 d_i \, a) - \frac{1}{n^2} \left(\sum_{i=1}^{n} d_i + n a \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} d_i^2 + \frac{1}{n} (n a^2) + \frac{2a}{n} \sum_{i=1}^{n} d_i - \frac{1}{n^2} \left[\left(\sum_{i=1}^{n} d_i \right)^2 + n^2 a^2 + 2 n a \sum_{i=1}^{n} d_i \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} d_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} d_i \right)^2$$

$$\therefore \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n} - \left(\frac{\sum_{i=1}^{n} d_i}{n}\right)^2} \qquad \ldots \text{(iv)}$$

In case of grouped data, this formula takes the shape

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i\, d_i^2}{n} - \left(\frac{\sum_{i=1}^{n} f_i\, d_i}{n}\right)^2}$$

where $n = \sum_{i=1}^{n} f_i$ . . . (v)

We can still modify this formula by defining $u_i = \frac{y_i - a}{c}$, c being the class size.

We are assuming that all classes are of equal width and the frequency of each class is centred at its class mark. In this method

$$\sigma = c\sqrt{\frac{\sum f_i\, \mu_i^2}{n} - \left(\frac{\sum_{i=1}^{n} f_i\, u_i}{n}\right)^2} \qquad \ldots \text{(vi)}$$

Standard deviation is usually abbreviated as SD.

Example 3.8

Find the SD of the first n natural numbers.

Solution

The series is 1, 2, 3, . . . , n.

We know that $$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left[\frac{\sum_{i=1}^{n} x_i}{n}\right]^2 \qquad \ldots \text{(i)}$$

$$\therefore \qquad \sigma = \sqrt{\frac{\sum n^2}{n} - \left(\frac{\sum n}{n}\right)^2} \qquad (\Theta\ x_i = i \text{ in this case})$$

$$= \sqrt{\frac{n\,(n+1)\,(2n+1)}{6n} - \frac{n^2\,(n+1)^2}{4n^2}}$$

$$= \sqrt{(n+1)\left(\frac{2n+1}{6} - \frac{n+1}{4}\right)}$$

$$= \sqrt{(n+1)\left(\frac{4n+2-3n-3}{12}\right)} = \sqrt{\frac{n^2-1}{12}}$$

Example 3.9

The scores of a batsman in 10 different matches were 38, 70, 48, 34, 42, 55, 63, 46, 54, 44. Find the MD and SD of these scores.

Solution

The scores arranged in ascending order are 34, 38, 42, 44, 46, 48, 54, 55, 63, 70.

Number of observations = 10

$\therefore$ Median = Mean of $\frac{10}{2}th$ and $\left(\frac{10}{2}+1\right)^{th}$ observations

= Mean of 5th and 6th observations

$= \frac{46+48}{2} = 47$.

x_i	34	38	42	44	46	48	54	55	63	70	Total
$\lvert x_i - 47 \rvert$	13	9	5	3	1	1	7	8	16	23	86

Hence $\text{MD} = \frac{\sum_{i=1}^{10} |x_i - \text{median}|}{10} = \frac{13+9+5+3+1+1+7+8+16+23}{10}$

$= \frac{86}{10} = 8.6$

To find the SD, let $a = 48$, consider the following table

x_i	$d_i = x_i - a$	d_i^2
34	– 14	196
38	– 10	100
42	– 6	36
44	– 4	16
46	– 2	4
48	0	0
54	6	36
55	7	49
63	15	225
70	22	484
Total	14	1146

$\therefore \quad \sigma^2 = \frac{\sum d_i^2}{n} - \left(\frac{\sum d_i}{n}\right)^2$

$= \frac{1146}{10} - \left(\frac{14}{10}\right)^2 = 114.6 - 1.96 = 112.64$

$\therefore \quad \sigma = \sqrt{112.64} = 10.61$ nearly.

Example 3.10

In a survey of 950 families in a village, the following distribution of number of children was obtained.

No. of Children	0 – 2	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12
No. of Families	272	328	205	120	15	10

Find the mean, median and the standard deviation.

Solution

Let us take $a = 5$ the assumed mean. We construct the following table

Class-interval	y_i	f_i	Comm. f_i	$u_i = \frac{y_i - 5}{2}$	u_i^2	$f_i u_i$	$f_i u_i^2$
0 – 2	1	272	272	– 2	4	– 544	1088
2 – 4	3	328	600	– 1	1	– 328	328
4 – 6	5	205	805	0	0	0	0
6 – 8	7	120	925	1	1	120	120
8 – 10	9	15	940	2	4	30	60
10 – 12	11	10	950	3	9	30	90
Total		950				– 692	1686

$$\text{Mean} = a + \frac{\Sigma f_i u_i}{\Sigma f_i} \times c = 5 + \frac{-692}{950} \times 2 = 5 - \frac{692}{475}$$

$$= 5 - 1.457 = 3.543$$

$$\text{SD} = c \times \sqrt{\frac{\Sigma f_i u_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i u_i}{\Sigma f_i}\right)^2}$$

$$= 2 \times \sqrt{\frac{1686}{950} - \left(\frac{-692}{950}\right)^2}$$

$$= 2 \times \sqrt{1.7747 - 0.5300} = 2 \times \sqrt{1.2442}$$

$$= 2 \times 1.115 = 2.23 \text{ nearly.}$$

To find median, we note $\frac{n}{2} = \frac{950}{2} = 475$, and the median class is 2 – 4.

$$\text{Hence median} = 2 + \frac{475 - 272}{328} \times 2 \left(l + \frac{\frac{n}{2} - c_{f-1}}{f_m} c_m \right)$$

$$= 2 + 1.238 = 3.238 \text{ nearly.}$$

Example 3.11

Calculate the mean deviation for the following data :

Marks	No. of Children
0 – 10	5
10 – 20	8
20 – 30	15
30 – 40	16
40 – 50	6

Solution

Construct the table (last two columns to be completed after the calculation of median).

Class	y_i	f_i	Comm. f_i	$\lvert y_i - 28\rvert$	$f_i \lvert y_i - 28\rvert$
0 – 10	5	5	5	23	115
10 – 20	15	8	13	13	104
20 – 30	25	15	28	3	45
30 – 40	35	16	44	7	112
40 – 50	45	6	50	17	102
Total		50			478

$$n = 50 \Rightarrow \frac{n}{2} = 25 .$$

The class of median is 20 – 30.

$$\text{Hence median } = 20 + \frac{25 - 13}{15} \times 10 \quad \Bigg| \quad l + \frac{\frac{n}{2} - c_{f-1}}{f_m} \times c_m$$

$$= 20 + 8 = 28$$

$$\text{M. D} = \frac{478}{50} = 9.56 \quad \Bigg| \quad \frac{\Sigma f_i \lvert y_i - 28\rvert}{\Sigma f_i}$$

Example 3.12

The score of 48 children in a test are shown in the following frequency table :

Score	71	76	79	83	86	89	92	97	101	103	107	110	114
Frequency	4	3	4	5	6	5	4	4	3	3	3	2	2

Find σ^2.

Solution

Let $a = 90$, the assumed mean. Construct the following table

x_i	f_i	$d_i = x_i - a$	d_i^2	$f_i d_i$	$f_i d_i^2$
71	4	– 19	361	– 76	1444
76	3	– 14	196	– 42	588
79	4	– 11	121	– 44	484
83	5	– 7	49	– 35	245
86	6	– 4	16	– 24	96
89	5	– 1	1	– 5	5
92	4	2	4	8	16
97	4	7	49	28	196
101	3	11	121	33	363

103	3	13	169	39	507
107	3	17	289	51	867
110	2	20	400	40	800
114	2	24	576	48	1152
Total	48			21	6763

$$\sigma^2 = \frac{6763}{48} - \left(\frac{21}{48}\right)^2 \quad \Big| \quad \frac{\Sigma f_i d_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i d_i}{\Sigma f_i}\right)^2$$

$$= 140.896 - 0.191 = 140.705 \text{ nearly.}$$

SAQ 1

(a) The postal expenses on the letters dispatched from an office on a given day is given in the following frequency distribution :

Postage (P)	15	30	35	60	70
Number of Letters	47	33	56	41	25

Find the mean postage per letter.

(b) The mean age (in years) per student and the number of students in each of the four classes of two primary schools are given below :

	School A		**School B**	
	No.	**Mean Age**	**No.**	**Mean Age**
Class I	6	6.2	25	7.1
Class II	10	7.5	32	8.4
Class III	28	8.6	12	9.2
Class IV	30	10.0	4	10.7

Obtain the mean age per student for the two schools.

(c) The measurements (in mm) of the diameters of the heads of 107 screws gave the following frequencies distribution.

Diameter	33 – 35	36 – 38	39 – 41	42 – 44	45 – 47
Frequency	17	19	23	21	27

Find the mean head diameter per screw.

(d) The marks obtained out of 50 by 102 students in a test were recorded and were according to the following frequency table :

Marks	20	22	23	24	26	31	38	43
Number of Students	8	15	28	27	20	2	1	1

Obtain the median and describe what information it conveys.

SAQ 2

(a) The following table gives the frequency distribution of married women by age at marriage :

Age (in years)	15 – 19	20 - 24	25 - 29	30 - 34	35 - 39
Frequency	53	140	98	32	12
Age (in years)	40 - 44	45 – 49	50 – 54	55 – 59	60 and above
Frequency	9	5	3	3	2

Calculate the median.

(b) The following table gives the weekly consumption of electricity of 50 families. Find the mean and median weekly consumption :

Weekly Consumption	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Number of Families	6	12	18	3	1

(c) The following data is about the number of days patients stayed in a hospital after an operation. Calculate the SD.

Hospital Stay (in days)	1 – 4	4 – 7	7 – 10	10 – 13	13 – 16	16 – 19	19 – 22
Number of Patients	32	108	67	28	14	7	3

(d) Calculate the SD for the following data :

Wages per Week up to (Rs.)	15	30	45	60	75	90	105	120
Number of Workers	12	30	65	107	157	202	220	225

SAQ 3

(a) Find the SD of the following table :

x_i	140	145	150	155	160	165	170	175
f_i	4	6	15	30	36	24	8	2

(b) The length (in cm) of 10 small pieces of cloth were :

5, 3, 9, 12, 3, 10, 12, 21, 18, 12

Find the mean deviation and the standard deviation.

(c) Find the mean deviation for the following observations

3, 3, 5, 9, 10, 12, 12, 12, 18, 21, 21.

3.6 DEVELOPMENT OF PROBABILITY THEORY

The word *probability* has several meanings in ordinary conversation. The most common one is the interpretation of probability as relative frequency, for which simple games involving coins, cards, dice, and roulette wheels provide examples. The distinctive feature of games of chance is that the outcome of a given trial cannot be predicted with certainty, although the collective results of a large number of trials display some regularity. For example, the statement that the probability of "heads" in tossing a coin equals one-half, according to the relative frequency interpretation, implies that in a large number of tosses, the relative frequency with which "head" actually occurs will be approximately one-half, although it contains no implication concerning the outcome of any given toss. There are many similar examples involving collections of people, molecules of a gas, genes, and so on. Actuarial statements about the life expectancy for persons of a certain age describe the collective experience of a large number of individuals, but do not purport to say what will happen to any particular person. Similarly, predictions about the chance of a genetic disease occurring in a child of parents having a known genetic makeup, are statements about relative frequencies of occurrence of disease in a large number of cases, but are not predictions about a given individual.

This unit contains a description of the important mathematical concepts of probability theory, illustrated by some of the applications that have stimulated their development. Since applications inevitably involve simplifying assumptions that focus on some features of a problem at the expense of others, it is advantageous to begin, like the 17th-century mathematicians mentioned above, by thinking about simple experiments, such as tossing a coin or rolling a dice, and later to see how these apparently frivolous investigations relate to important scientific questions.

3.6.1 asic Concepts of Probability

The word "Probability" or "Chance" is very commonly used in day-to-day conversation. For example, read the statements given below

- Probability that it will rain tomorrow, is 50 percent
- The chance of passing the examination, is 90 percent
- There is a 80-20 chance that India will win the next one-day international cricket match against New Zealand.

There is always some uncertainty about the happening of the events concerning mass phenomena In this unit, we shall introduce some elementary probability concepts, indicate how probabilities can be interpreted and then show how rules of probability can be applied to compute the probability of combinations of events. The methodologies of computing probability will then permit us to express, in precise language, such informed statements as the ones given above.

If an event can occur in N_A outcomes out of a total of N possible and equally likely outcome, then the probability that the event will occur is

$$P(A) = \frac{N_A}{N} \qquad \ldots (11.1)$$

where $P(A)$ = Probability of an event occurring,

N_A = Number of successful outcomes, and

N = Total number of possible outcomes.

This definition can be used when the number of outcomes is known or when the number of outcomes is found by experimentation.

Now let us try to compute the probability of some simple problem.

Example 3.13

A box contains six 20 Ω resistors and fifteen 40 Ω resistors. The resistors are all unmarked and of the same physical size. If one resistor is picked out at random, determine the probability of its resistance being 20 Ω.

Solution

Let, A be the set of all 20 Ω resistors and B be the set of all 40 Ω resistors.

Now, $N = 6 + 15 = 21$, and $N_A = 6$

Then, $P(A) = \frac{N_A}{N} = \frac{6}{21} = \frac{2}{7}$

3.6.2 Some Useful Terms

Sample Space and Events

The concept of probability is relevant to experiments that have somewhat uncertain outcomes. These are the situation in which, despite every effort to maintain fixed conditions, some variation of the result in repeated trials of the experiments, is unavoidable. In probability, the term experiment is not restricted to laboratory experiment but includes any activity that results in the collection of data pertaining to phenomena that exhibit variation.

Let us begin study of probability by introducing the **concept of sample space and events**.

The collection of all possible distinct outcomes of an experiment, is called the *sample space* of outcomes; each of the distinct outcomes is called a *simple event*, an elementary outcome or an element of the sample space. The sample space is usually denoted by S.

Events

In our study of probability, we will be interested not only in the individual outcomes from S but also in any collection of outcomes from S.

An event is any collection (subject) of outcomes contained in the sample space. An event is said to be *simple* if it consists of exactly one outcome and compound if it consists of more than one outcome.

A sample space consisting of either a finite or a countably infinite number of element is called a *discrete sample space*. When the sample space includes all the numbers in some interval (finite or infinite) of the real line, it is called *continuous sample space*.

Random Experiment

If in each trial of an experiment repeated under identical conditions the outcome is not unique, but may be any of the possible outcomes, then such an experiment is known as random experiment.

For example, tossing a coin, throwing a die, drawing a card from a pack of cards are all examples of random experiment.

Exhaustive Events

The set of all possible outcomes of a single performance of an experiment is called exhaustive events. Each outcome is called a sample point. In case of tossing a coin once, $S = \{H, T\}$.

Mutually Exclusive Events

Two events are said to be mutually exclusive if the occurrence of one event makes the other event impossible. Thus, if on one throw of a die, a 3 occurs (event A), the event B, say a 5, cannot occur.

Some Concepts from Set Theory

An event is nothing but a set of outcomes, so that concepts and results from elementary set theory can be used to study events. The following concepts from set theory will be helpful for future developments.

(i) The union of two events A and B, denoted by $A \cup B$, is the event consisting of all outcomes that are either in A or in B or in both events A and B.

(ii) The intersection of two events A and B, denoted by $A \cap B$, is the event consisting of all outcomes that are in both A and B.

(iii) The complement of an event A, denoted by A^C, is the set of all outcomes in S that are not contained in A.

Example 3.14

Suppose a six-sided die is tossed once, and the outcome is the number on the up turned face

Let $A = \{1, 3, 5, 6\}$, $B = \{4, 5, 6\}$ and $C = \{1, 2\}$

Then $A \cup B = \{1, 3, 4, 5, 6\}$

$A \cup C = \{1, 2, 3, 5, 6\}$ also $A \cap C = \{1\}$

$A^C = \{2, 4\}$, $B^C = \{1, 2, 3\}$

Note : When two events have no outcomes in common they are said to be *mutually exclusive* or *disjoint events*. You can verify that in the above example, B and C have no elements in common, so B and C are mutually exclusive or disjoint events.

3.7 THEOREMS OF PROBABILITY

As the discussion of the subject matter is from application/engineering point of view, theorems are only stated, without any proofs.

Theorem 1

Probability is expressed as a number between 0 and 1.0, where a value of 1.0 denotes a certainty that an event will occur and a value of 0 denotes a certainty that an event will not occur.

Theorem 2

If $P(A)$ is the probability that event A will occur, then the probability that A will not occur is $1 - P(A)$.

Example 3.15

If the probability of obtaining a defective electrical heater is 0.08, what is the probability of obtaining an acceptable electrical heater?

Solution

The required probability is $= 1 - P(A)$

$$= 1 - 0.08 = 0.92.$$

Therefore, the probability of obtaining an acceptable electrical heater is 0.92.

Theorem 3

If *A* and *B* are two mutually exclusive events, then the probability that either event *A* or event *B* will occur is the sum of their respective probabilities, i.e.

$$P(A \text{ or } B) = P(A) + P(B) \qquad \ldots (3.14)$$

Whenever an 'or' is verbalized, the mathematical operation is addition. Theorem 3 is stated for two events – it is equally applicable for more than two events. Thus

$$P(A \text{ or } B \text{ or } \ldots \text{ or } F) = P(A) + P(B) + \ldots + P(F),$$

Provided A, B, . . . , and F are mutually exclusive.

Example 3.16

If the 291 parts, described in Table 11.1, are contained in a box, then

Table 11.1 : Inspection Results by Supplier

Supplier	Number Acceptable	Number Defective	Total
A	58	5	63
B	133	8	141
C	83	4	87
Total	274	17	291

(i) what is the probability of selecting a random part produced by supplier A or by supplier C?

(ii) what is the probability of selecting a defective part from supplier A or an acceptable part from supplier C?

(iii) what is the probability that a randomly selected part will be from supplier C, a defective from supplier A or an acceptable from supplier B?

Solution

Now the event that a part is by supplier A is mutually disjoint from the event that a part is supplied by supplier B. Similarly, for (A and C) and (B and C). Therefore,

(i) $P(A \text{ or } C) = P(A) + P(C) = \dfrac{63}{291} + \dfrac{87}{291} = 0.515$

(ii) P (defective A or acceptable C)

$= P$ (defective A) + P (acceptable C)

$$= \frac{5}{291} + \frac{83}{291} = 0.302$$

(iii) P (C, or defective A, or acceptable B)

$= P(C) + P$ (defective A) + P (acceptable B)

$$= \frac{87}{291} + \frac{5}{291} + \frac{133}{291} = 0.7732$$

Theorem 4

If event A and event B are not mutually exclusive events, then the probability of either event A or event B or both is given by

$$\boldsymbol{P\ (A \text{ or } B) = P\ (A) + P\ (B) - P\ (\text{both})} \qquad \ldots (3.15)$$

Events that are not mutually exclusive have some outcome in common.

Example 3.17

If the 291 parts, described in Table 11.1, are contained in a box, what is the probability that a randomly selected part will be from supplier A or a defective?

Solution

$$P\ (A \text{ or defective}) = P\ (A) + P\ (\text{defective}) - P\ (A \text{ and defective})$$

$$= \frac{63}{291} + \frac{17}{291} - \frac{5}{291} = 0.2577$$

Theorem 5

The sum of the probabilities of all the events of a situation is equal to 1.00

$$\boldsymbol{P\ (A) + P\ (B) + \ldots + P\ (N) = 1.00} \qquad \ldots (3.16)$$

where *A*, *B*, . . . , *N* are all possible events of a situation.

Independent Events

A pair of events are said to be independent if occurrence of one has no influence on the probability of occurrence the other event. Further, we may generalise the concept of independent events to *n* events.

Theorem 6

If A and B are independent events, then the probability of both A and B occurring is the product of their respective probabilities.

$$\boldsymbol{P\ (A \text{ and } B) = P\ (A) \times P\ (B)} \qquad \ldots (3.17)$$

This theorem is also known as the multiplicative law of probabilities.

Example 3.18

If the 291 parts, described in Table 11.1, are contained in a box, what is the probability that two randomly selected parts will be from supplier B and C? Assume that the first part is returned to the box before second part is selected.

Solution

If the earlier selected part is returned to the box before the next selection is made, then any two selections are independent of each other.

$$P\ (B \text{ and } C) = P\ (B) \times P\ (C)$$

$$= \frac{141}{291} \times \frac{87}{291} = 0.145$$

A dependent event is one whose occurrence influences the probability of the other event or events.

Theorem 7

If A and B are dependent events, the probability of both A and B occurring is the product of the probability of A and the probability of B, provided event A has occurred, i.e.

$$\boldsymbol{P\ (A \text{ and } B) = P\ (A) \times P\ (B / A)} \qquad \ldots (3.18)$$

The symbol $P(B / A)$ denotes the probability of event B provided that event A has occurred. This theorem is known as Conditional Probability Theorem, since the probability of the second event depends on the result of the first event.

Example 3.19

Let us consider the previous problem with the assumption that the first part was not returned to the box before the second part is selected. What is the probability of B and C in two successive selections?

Solution

$$P(B \text{ and } C) = P(B) \times P(C / B)$$

$$= \frac{141}{291} \times \frac{87}{290} = 0.1454$$

Example 3.20

Let us consider the previous problem.

What is the probability of both parts from supplier A?

Solution

$$P(A \text{ and } A) = P(A) \times P(A / A)$$

$$= \frac{63}{291} \times \frac{62}{290} = 0.046$$

Permutation and Combination

While finding the probabilities of an event, the concepts of permutation and combination and various results concerning the concepts, are often required.

Permutation

A *permutation* of a number of objects is their arrangement in some definite order. The permutations of the word "MAD", are MAD, MDA, ADM, AMD, DAM, DMA. In this simple example, there are 3 objects in the set, and we arranged them in groups of 3 to obtain six permutation. This is referred to as a permutation of n objects taking r at a time, where $n = 3$ and $r = 3$. Let us consider the case when there are 4 objects and 2 objects are taken at a time. Using the four letters of the word "SHIV" to represent the four objects, the permutations are SH, SI, SV, HS, IS, VS, HI, HV, IH, VH, IV, VI. As the number of objects n and the number that are taken at one time, r, become larger, it becomes a difficult task to list all the permutations.

The formula to find the number of permutation more easily is

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

where $^{n}P_{r}$ = number of permutations of n objects taken r of them at a time,

n = total number of objects, and

r = number of objects selected out of the total number.

Example 3.21

How many permutation are there of 9 objects taken 5 at a time?

Solution

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

$$= \frac{9!}{(9-5)!} = \frac{9!}{4!} = \frac{9.8.7.6.5.4!}{4!} = 15120$$

Combinations

If the way in which the objects are ordered is unimportant, then we have a combination. The word "MAD" has six permutations when the 3 objects are taken 3 at a time. However, there is only one combination, since the same three letters are in a different order. The word "SHIV" has 12 permutations when the 4 letters are taken 2, at a time, but the number of combination is SH, SI, SV, HI, HV, IV which gives a total of six.

The number of combinations of n different object taken r at a time is denoted by ${}^nC_r$. If we take any one of the combinations, its r objects can be arranged in r ! ways. So the total number of arrangements which can be obtained from all the combinations is

$${}^nP_r = {}^nC_r \; r!$$

$$\Rightarrow \qquad {}^nC_r = \frac{n!}{(n-r)!\,r!}$$

You may verify that ${}^nC_r = {}^nC_{n-r}$

where ${}^nC_r$ = number of combinations of n objects taken r at a time,

n = total number of objects, and

r = number of objects selected out of the total number.

Example 3.22

If there are 10 balls in a box, how many different combinations are possible if 6 are selected at a time.

Solution

We know that

$${}^nC_r = \frac{n!}{(n-r)!\,r!}$$

Here, $${}^{10}C_6 = \frac{10!}{(10-6)!\,6!}$$

$$= \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}$$

$$= 210.$$

Example 3.23

A bag contains four white and two black balls and a second bag contains three of each colour. A bag is selected at random, and a ball is then drawn at random from the bag chosen. What is the probability that the ball is white?

Solution

The required probability $= \frac{1}{2}\left(\frac{4}{6} + \frac{3}{6}\right)$

$$= \frac{7}{12}$$

Example 3.24

An article manufactured by a company consists of two parts A and B. In the process of manufacture of part A, 9 out of 100 are likely to be defective. Similarly, 5 out of 100 are likely to be defective in the manufacture of part B. Calculate the probability that the assembled article will not be defective (assuming that the events of finding the part A non-defective and that of B are independent).

Solution

The required probability = 0.91×0.95

$= 0.8645$

Example 3.25

A bag contains 10 white and 15 black balls. Two balls are drawn in succession (i.e. without replacement). What is the probability that first is white and second is black?

Solution

The required probability $= \frac{10}{25} \times \frac{15}{24}$

$= 0.25$

SAQ 4

(a) A part-time student is taking two courses namely, Statistics and Finance. The probability that the student will pass the statistics course is 0.60 and the probability of passing the finance course is 0.70. Find the probability that the student

(i) will pass at least one course

(ii) will fail both courses

(b) A machine contains a component C that is vital to its operation. The reliability of component C is 80%. To improve the reliability of a machine, a similar component is used in parallel to form a system. The machine will work provided that one of these components functions correctly. Calculate the reliability of the system S.

(c) A class consists of 80 students, 25 of them girls and 55 boys. While 10 of them are rich and the remaining poor, it is found that 20 are fair complexioned. What is the probability of selecting a fair complexioned rich girl or a poor boy who is not fair complexioned?

(d) A speak the truth in 70% cases and B speaks the truth in 80% cases. What is the probability that they will say the same thing while describing a single event?

(e) The chances that doctor A will diagnose a disease X is 60%. The chances that a patient will die by his treatment after correct diagnosis is 40% and the chances of death by wrong diagnosis is 70%. A patient of doctor A who had disease X, died. What is the chance that his disease was diagnosed correctly?

(f) An anti-aircraft gun can take a maximum of four shots on enemy's plane moving away from it. The probabilities of hitting the plane at first, second, third and fourth shots are 0.4, 0.3, 0.2 and 0.1 respectively. Find the probability that the gun hits the plane.

3.8 THE PRINCIPLE OF ADDITIVITY

The last section illustrates the fundamental principle that, if the event whose probability is sought can be represented as the union of several other events that have no outcomes in common ("at most one head" is the union of "no heads" and "exactly one head"), then the probability of the union is the sum of the probabilities of the individual events making up the union. To describe this situation symbolically, let S denotes the sample space. For two events A and B, the intersection of A and B is the set of all experimental outcomes belonging to both A and B and is denoted by $A \cap B$; the union of A and B is the set of all experimental outcomes belonging to A or B (or both) and is denoted by $A \cup B$. The impossible event – the event containing no outcomes – is denoted by ϕ. The probability of an event A is written as $P(A)$.

The principle of addition of probabilities is that, if $A_1, A_2, \ldots, A_n$ are events with $A_i \cap A_j = \phi$ for all pairs $i \neq j$, then

$$P(A_1 \cup A_2 \cup \ldots A_n) = P(A_1) + P(A_2) + \ldots + P(A_n) \qquad \ldots (3.19)$$

Eq. (3.19) is consistent with the relative frequency interpretation of probabilities; for, if $A_i \cap A_j = \phi$ for all $i \neq j$, the relative frequency with which at least one of the A_i occurs equals the sum of the relative frequencies with which the individual A_i occur.

An elementary, useful consequence of Eq. (3.19) is the following. With each event A is associated the complementary event A^c consisting of those experimental outcomes that do not belong to A. Since $A \cap A^c = \phi$, $A \cup A^c = S$, and $P(S) = 1$ (where S denotes the sample space), it follows from (3.19) that $P(A^c) = 1 - P(A)$. For example, the probability of "at least one head" in n tosses of a coin is one minus the probability of "no head," or $1 - 1/2^n$.

3.9 CONDITIONAL PROBABILITY

Suppose two balls are drawn sequentially without replacement from an urn containing r red and b black balls. The probability of getting a red ball on the first draw is $r / (r + b)$. If, however, one is told that a red ball was obtained on the first draw, the conditional probability of getting a red ball on the second draw is $(r - 1) / (r + b - 1)$, because for the second draw there are $r + b - 1$ balls in the urn, of which $r - 1$ are red. Similarly, if one is told that the first ball drawn is black, the conditional probability of getting red on the second draw is $r / (r + b - 1)$.

In a number of trials, the relative frequency with which B occurs among those trials in which A occurs is just the frequency of occurrence of $A \cap B$ divided by the frequency of occurrence of A. This suggests that the conditional probability of B given A (denoted by $P(B / A)$) should be defined by

$$P(B / A) = \frac{P(A \cap B)}{P(A)} \qquad \ldots (3.22)$$

If A denotes a red ball on the first draw and B a red ball on the second draw in the experiment of the preceding paragraph, then $P(A) = r / (r + b)$ and

$$P(A \cap B) = \frac{r(r-1)}{[(r+b)(r+b-1)]} \qquad \ldots (3.23)$$

$\therefore$ Eq. (3.22) is consistent with the "obvious" answer derived above.

Rewriting Eq. (3.22) as $P(A \cap B) = P(A)\, P(B \mid A)$ and adding to this expression the same expression with A replaced by A^c ("not A") leads to the equality

$$P(B) = P(A \cap B) + P(A^c \cap B)$$

$$= P(A)\ P(B/A) + P(A^c)\ P(B/A^c) \qquad \ldots (3.24)$$

More generally, if $A_1, A_2, \ldots, A_n$ are mutually exclusive events so that exactly one of the A_k must occur and exhaustive, i.e. their union is the entire sample space. Essentially the same argument gives a fundamental relation, which is frequently called the **law of total probability**.

$$P\ (B) = P(A_1)\ P(B/A_1) + P(A_2)\ P(B/A_2) + \ldots + P(A_n)\ P(B/A_n) \qquad \ldots (3,25)$$

Independence

One of the most important concepts in probability theory is that of "independence". We discussed independence of events earlier also. Here, we discuss the concept from another perspective. The events A and B are said to be (stochastically) independent if $P\ (B \mid A) = P\ (B)$, or equivalently by Eq. (3.22) if

$$P(A \cap B) = P(A)\ P(B) \qquad \ldots (3.26)$$

The intuitive meaning of the definition in terms of conditional probabilities is that the probability of B is not changed by knowing that A has occurred. Eq. (3.26) shows that the definition is symmetric in A and B.

It is intuitively clear that, in drawing two balls with replacement from an urn containing r red and b black balls, the event "red ball on the first draw" and the event "red ball on the second draw" are independent. (This statement presupposes that the balls are thoroughly mixed before each draw.) An analysis of the $(r + b)^2$ equally likely outcomes of the experiment shows that relation (3.26) is indeed satisfied.

Example 3.26

A problem of statistics is given to three students A, B and C whose chances of solving it are $\frac{1}{2}, \frac{1}{3}$ and $\frac{1}{4}$ respectively. What is the probability that the problem will be solved?

Solution

Let A, B, C be the events of three students respectively solving the problem.

So, $\quad P(A) = \frac{1}{2},\ P(B) = \frac{1}{3},\ P(C) = \frac{1}{4}$.

Obviously, A, B, C are independent events.

Probability that the problem is not solved = $P(\overline{A})\ .\ P(\overline{B})\ .\ P(\overline{C})$

$$= \left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{3}\right)\left(1 - \frac{1}{4}\right)$$

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

The probability that the problem is solved $= 1 - \frac{1}{4} = \frac{3}{4}$

Example 3.27

If events A and B are independent and $P\ (A) = 0.15$, $P\ (A \cup B) = 0.45$, then find $P\ (B)$.

Solution

Since events A and B are independent,

$$P\ (A \cap B) = P\ (A)\ P\ (B)$$

So, $\quad P\ (A \cap B) = 0.15\ P\ (B)$

Now we know,

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

So, $\quad 0.45 = 0.15 + P(B) - 0.15\, P(B)$

$\Rightarrow \quad P(B) = 0.3529.$

SAQ 5

(a) A lot contains 20 articles. The probability that the lot contains exactly 2 defective articles is 0.4 and that it contains exactly 3 defective articles is 0.6. Articles are drawn from the lot at random one by one without replacement and tested till all the defective articles are found. What is the probability that this procedure ends at the twelfth testing?

(b) A box contains 2 black, 4 white and 3 red balls. One ball is drawn at a time randomly from the box till all the balls are drawn from it. Find the probability that the balls drawn are in the sequence of 2 black, 4 white and 3 red.

(c) An urn contains 5 balls. Two balls are drawn and found to be white. What is the probability that all the balls are white?

(d) An urn contains 10 white and 15 red balls. Two balls are drawn at random in succession. Find the probabilities of two balls together by virtue of the sampling

(i) with replacement, and

(ii) without replacement

3.10 BAYES'S THEOREM

Consider now the defining Eq. (3.22) for the conditional probability $P(E_n / A)$, where the E_i are mutually exclusive and, exhaustive. Substitution of $P(E_n)\, P(A/E_n)$ in the numerator of Eq. (3.22) and substitution of the right-hand side of the law of total probability in the denominator yields a result known as **Bayes's theorem** (after the 18th-century English clergyman Thomas Bayes) **or the law of inverse probability**; which is stated below :

Statement

"If $E_1, E_2, E_3, \ldots, E_n$ are n mutually exclusive and exhaustive events such that $P(E_i) > 0$ for each i and A is an arbitrary event for which $P(A) \neq 0$ then for each i, $1 \leq i \leq n$, the conditional probability of occurrence of E_i ($i = 1, 2, 3, \ldots, n$) given that A has occurred is given by

$$P(E_i / A) = \frac{P(E_i)\, P(A/E_i)}{P(E_1)\, P(A/E_1) + P(E_2)\, P(A/E_2) + \ldots + P(E_n) P(A/E_n)}$$

$$= \frac{P(E_i)\, P(A/E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A/E_i)}, \quad i = 1, 2, 3, \ldots, n\text{"}$$

Thus, Bayes's theorem connects the conditional probability $P(E_i / A)$ of E_i when A has occurred with the conditional probability $P(A / E_i)$ of A when E_i has occurred for each i and the probabilities of E_i's themselves.

The probabilities $P(E_1), P(E_2), \ldots, P(E_n)$ which are already given or known before conducting an experiment are termed as **a priori** probabilities and the conditional

probabilities $P(E_1/A), P(E_2/A), \ldots, P(E_n/A)$, which are computed after conducting the experiment (viz., occurrence of A) are termed as **posteriori** probabilities.

We shall now give proof of Bayes's Theorem.

Proof of Bayes's Theorem

Events $E_1, E_2, E_3, \ldots, E_n$ are mutually exclusive

$\therefore \quad E_1 \cup E_2 \cup E_3 \cup \ldots \cup E_n = S$

$\therefore \quad A = A \cap S = (A \cap E_1) \cup (A \cap E_2) \cup (A \cap E_3) \cup \ldots \cup (A \cap E_n)$

where $A \cap E_1, A \cap E_2, A \cap E_3, \ldots, A \cap E_n$ are all mutually exclusive events.

Hence by addition law of probability, we get

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3) + \ldots + P(A \cap E_n)$$

$$= P(E_1) \,.\, P(A/E_1) + P(E_2) \,.\, P(A/E_2) + P(E_3) \,.\, P(A/E_3) + \ldots + P(E_n) \,.\, P(A/E_n)$$

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

$\therefore$ For a fixed i,

$$P(E_i/A) = \frac{P(E_i \cap A)}{P(A)} = \frac{P(E_i) \,.\, P(A/E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A/E_i)}$$

$$= \frac{P(E_i) \,.\, P(A/E_i)}{P(E_1) \,.\, P(A/E_1) + P(E_2) \,.\, P(A/E_2) + \ldots + P(E_n) \,.\, P(A/E_n)}$$

where $\quad i = 1, 2, 3, \ldots, n$

We now consider some examples.

Example 3.28

Three urns A, B, C contain 6 red and 4 black balls, 2 red and 6 black balls and 1 red and 8 black balls respectively. An urn is chosen at random and a ball is drawn from the urn. If the ball drawn is red, find the probability that the ball was drawn from urn A.

Solution

Let E_1, E_2, E_3 denote the events of choosing urn A, B and C respectively. Let E denotes the event that the ball drawn is red. Then

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

$$P(E/E_1) = \frac{6}{10} = \frac{3}{5},\ P(E/E_2) = \frac{2}{8} = \frac{1}{4} \text{ and } P(E/E_3) = \frac{1}{9}$$

$$\therefore \quad P(E_1 \cap E) = \frac{1}{3} \times \frac{3}{5} = \frac{1}{5},$$

$$P(E_2 \cap E) = \frac{1}{3} \times \frac{1}{4} = \frac{1}{12}$$

and $\quad P(E_3 \cap E) = \frac{1}{3} \times \frac{1}{9} = \frac{1}{27}$

$$\therefore \quad P(E) = P(E_1 \cap E) + P(E_2 \cap E) + P(E_3 \cap E)$$

$$= P(E_1) \,.\, P(E/E_1) + P(E_2) \,.\, P(E/E_2) + P(E_3) \,.\, P(E/E_3)$$

$$= \frac{1}{3} \times \frac{3}{5} + \frac{1}{3} \times \frac{1}{4} + \frac{1}{3} \times \frac{1}{9}$$

$$= \frac{1}{5} + \frac{1}{12} + \frac{1}{27}$$

$$= \frac{108 + 45 + 20}{540} = \frac{173}{540}$$

$\therefore$ Required probability $P(E_1/E) = \dfrac{P(E_1 \cap E)}{P(E)}$

$$= \frac{\frac{1}{5}}{\frac{173}{540}} = \frac{108}{173}$$

SAQ 6

(a) In a bolt factory, machines A, B and C manufacture 25%, 35% and 40% of the total output respectively. Of their outputs, 5%, 4% and 2% are defective bolts. A bolt is chosen at random and found to be defective. What will be the probability that the bolt came from machine A, B and C?

(b) A factory manufacturing televisions has four units A, B, C and D. The units A, B, C, D manufacture 15%, 20%, 30% and 35% of the total output respectively. It was found that out of their output 1%, 2%, 2% and 3% are defective. A television is chosen at random from the total output, and found to be defective. What is the probability that it came from unit D?

3.11 RANDOM VARIABLES

Usually it is more convenient to associate numerical values with the outcomes of an experiment than to work directly with a non-numerical description such as "red ball on the first draw". For example, an outcome of the experiment of drawing n balls with replacement from an urn containing black and red balls is an n-tuple each component of which tells us whether a red or a black ball was drawn on the corresponding draw. This n-tuple may be conveniently represented by an n-tuple of ones and zeros, where the appearance of a one in the kth position indicates that a red ball was drawn on the kth draw. A quantity of particular interest is the number of red balls drawn, which is just the sum of the entries in this numerical description of the experimental outcome. Mathematically, a rule that associates a unique real number, with every element of a given set, is called a "real-valued function". In the history of statistics and probability, real-valued functions defined on a sample space have traditionally been called "**random variables**". Thus, if a sample space S has the generic element e, the outcome of an experiment, then a random variable is a real-valued function $X = X(e)$. Customarily, one omits the argument e in the notation for a random variable. For the experiment of drawing balls from an urn containing black and red balls, R, the number of red balls drawn, is a random variable. A particularly useful random variable is $1[A]$, the **indicator variable** of the event A, which equals 1 if A occurs and 0 otherwise. If A_k denotes the event "a red ball is drawn on the kth draw", then R has the useful representation : $R = 1[A_1] + \cdots + 1[A_n]$. A *constant* is a trivial random variable that always takes the same value regardless of the outcome of the experiment.

3.12 DISTRIBUTION FUNCTION

A random variable is a numerical description of the outcome of a statistical experiment. A random variable that may assume only a finite number or an infinite sequence of values is said to be *discrete*; one that may assume any value in some interval on the real number line is said to be *continuous*. For instance, a random variable representing the number of automobiles sold at a particular dealership on one day would be discrete, while a random variable representing the weight of a person in kilograms (or pounds) would be continuous.

The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable. For a discrete random variable x, the probability distribution is defined by a *probability mass function*, denoted by $f(x)$. This function provides the probability for each value of the random variable. In the development of the probability function for a discrete random variable, two conditions must be satisfied :

(i) $f(x)$ must be nonnegative for each value of the random variable, and

(ii) the sum of the probabilities, one for each value of the random variable, must equal one.

A continuous random variable may assume any value in an interval on the real number line or in a collection of intervals. Since there is an infinite number of values in any interval, it is not meaningful to talk about the probability that the random variable will take on a specific value; instead, the probability that a continuous random variable *will lie* within a given interval is considered.

In the continuous case, the counterpart of the probability mass function is the *probability density function*, also denoted by $f(x)$. For a continuous random variable, the probability density function provides the height or value of the function at any particular value of x; it does not directly give the probability of the random variable taking on a specific value. However, the area under the graph of $f(x)$ corresponding to some interval, obtained by computing the integral of $f(x)$ over that interval, provides the probability that the variable will take on a value within that interval. A probability density function must satisfy two requirements :

(i) $f(x)$ must be nonnegative for each value of the random variable, and

(ii) the integral over all values of the random variable must equal one.

The expected value, or mean, of a random variable – denoted by E (x) or μ – is a weighted average of the values the random variable may assume. In the discrete case, the weights are given by the probability mass function, and in the continuous case the weights are given by the probability density function. The formulas for computing the expected values of discrete and continuous random variables are given by Eqs. (3.27) and (3.28), respectively.

$$E(x) = \Sigma\, x f(x) \qquad \ldots (3.27)$$

$$E(x) = \int x f(x)\, dx \qquad \ldots (3.28)$$

The *variance of a random variable, denoted by Var (x) or* σ^2, is a weighted average of the squared deviations from the mean. In the discrete case, the weights are given by the probability mass function, and in the continuous case, the weights are given by the probability density function. The formulas for computing the variances of discrete and continuous random variables are given by Eqs. (3.29) and (3.30), respectively. The standard deviation, denoted by σ, is the positive square root of the variance. Since the standard deviation is measured in the same units as the random variable and the variance is measured in squared units, the standard deviation is often the preferred measure.

$$\text{Var}(x) = \sigma^2 = \Sigma\, (x - \mu)^2 f(x) \qquad \ldots (3.29)$$

$$\text{Var}(x) = \sigma^2 = \int (x - \mu)^2 f(x)\, dx \qquad \ldots (3.30)$$

Probability Distribution

If an experiment is performed on a random variable X and it takes value a, then the probability of this event is denoted by $P(X = a)$. Suppose the event is $X \leq a$, the corresponding probability is denoted by $P(X \leq a)$. For the event $X > a$, the probability is $P(X > a)$. These two events are mutually exclusive because if the event $X > a$ occurs, the other cannot. Hence,

$$P(X \leq a) + P(X > a) = 1. \qquad \ldots (3.31)$$

Recalling, the already defined random variable X is said to be discrete if it can take at most a countable number of values denoted by x_i, $i = 1, N$. Here, N is a number, which can tend to infinity.

Let $x_1, x_2 \ldots\ldots$ be the values for each of which X has a positive probability. Let p_i be probability of x_i. Then a probability function of X can be defined for each of such x as

$$f(x) = \begin{cases} p_j & \text{where } x = x_j,\ j = 1, 2, \ldots, N \\ 0 & \text{otherwise} \end{cases} \qquad \ldots (3.32)$$

If the probability function is known, then for any unique number the probability can be determined. This is called the *probability function* of the random variable X. For any a and b ($>a$), the probability that $a < X \leq b$ is

$$P(a < X \leq b) = P(X \leq \text{b}) - P(X < a) = F(b) - F(a) \qquad \ldots (3.33)$$

$P(X \leq x)$ depends on the choice of x, i.e. it is a function of x, which is called the *distribution function* of X and is denoted by $F(x)$.

The measures associated with a probability distribution are *expected value* and variance. The expected value is the weighted average of the values of the random variable. It is also called central tendency. The formula for calculating the average is :

$$E(x) = \sum_{i}^{n} x_i P(x_i) = x_1 P(x_1) + x_2 P(x_2) + \ldots + x_n P(x_n) \qquad \ldots (3.34)$$

The variance represents the variability in distribution. It is calculated as :

$$\text{Variance} = \sum_{i}^{n} [x_i - E(x)]^2 \ P(x_i) \qquad \ldots (3.35)$$

where $\left[x_i - E(x)\right]$ is the difference between each value of the random variable and the expected value.

Any probability distribution is described about the mean. If the distribution is symmetrical about the mean it is termed normal distribution. In the next three sub-sections, three important theoretical probability mass/density functions shall be introduced. These functions are used to describe discrete and continuous random variables. The use of continuous random variables is shown through a project concerning the high temperature during a given day etc.

3.12.1 The Binomial Distribution

Two of the most widely used discrete probability distributions are the Binomial and Poisson.

A binomial experiment has four properties :

(i) it consists of a sequence of n identical trials,

(ii) exactly one of the two outcomes viz., success or failure, is possible on each trial,

(iii) the probability of success on any trial, denoted by p, does not change from trial to trial, and

(iv) the trials are independent.

To Find the Probability of the Happening of an Event Once, Twice, Thrice, . . . *r* Times . . . Exactly in *n* Trials.

Let the probability of the happening of an event A in one trial be p and its probability of not happening be $1 - p = q$.

We assume that there are n trials and the event A happens r times and fails to happen $n - r$ times.

This may be shown as follows :

$$\underbrace{A\,A\,A\cdots A}_{r \text{ times}} \quad \underbrace{\bar{A}\,\bar{A}\,\bar{A}\cdots \bar{A}}_{n-r \text{ times}} \qquad \ldots (3.36)$$

A indicates its happening, $\bar{A}$ its failure and $P(A) = p$ and $P(\bar{A}) = q$.

We see that Eq. (3.36) has the probability

$$\underbrace{p\,p\,p\cdots p}_{r \text{ times}} \cdot \underbrace{q\,q\,q\cdots q}_{n-r \text{ times}} = p^r \cdot q^{n-r} \qquad \ldots (3.37)$$

Clearly representation (3.36) is merely one way of arranging r A' s and $(n - r)$ $\bar{A}$'s

The probability of (3.36) $= p^r \cdot q^{n-r} \times$ Number of different arrangements of

r A's and $(n - r)$ $\bar{A}$'s,

The number of different arrangement of r A's and $(n - r)$ $\bar{A}$'s $= {}^nC_r$

$\therefore$ Probability of the happening of an event r times $= {}^nC_r \; p^r \; q^{n-r}$

$$P(r) = {}^nC_r \; p^r \; q^{n-r} \qquad (r = 0, 1, 2, \ldots, n) \qquad \ldots (3.38)$$

$$= (r + 1)^{\text{th}} \text{ term of Binomial expansion of } (q + p)^n$$

For instance, suppose that it is known that 10 percent of the owners of two-year old automobiles have had problems with their automobile's electrical system. To compute the probability of finding exactly 2 owners that have had electrical system problems out of a group of 10 owners, the binomial probability mass function can be used by setting $n = 10$, $r = 2$, and $p = 0.1$ in Eq. (3.38); for this case, the probability is 0.1937.

Note that

If $r = 0$, probability of happening of an event 0 times $= {}^nC_0 \; q^n \; p^0 = q^n$

If $r = 1$, probability of happening of an event 1 time $= {}^nC_1 \; q^{n-1} \; p$

If $r = 2$, probability of happening of an event 2 times $= {}^nC_2 \; q^{n-2} \; p^2$

If $r = 3$, probability of happening of an event 3 times $= {}^nC_3 \; q^{n-3} \; p^3$ and so on.

These terms are clearly the successive terms in the Binomial expansion of $(q + p)^n$.

Hence it is called the Binomial Distribution.

Example 3.29

If on an average one ship in every ten is wrecked, find the probability that out of 5 ships expected to arrive, 4 at least will arrive safely.

Solution

Out of 10 ships, one ship is wrecked, i.e. nine ships out of ten ships are safe.

$$p \text{ (safety)} = \frac{9}{10}$$

P (at least 4 ships out of 5 are safe) $= P$ (4 or 5)

$$= P(4) + P(5)$$

$$= {}^5C_4\ p^4 q^{5-4} + {}^5C_5\ p^5 q^0$$

$$= 5\left(\frac{9}{10}\right)^4 \left(\frac{1}{10}\right) + \left(\frac{9}{10}\right)^5$$

$$= \left(\frac{9}{10}\right)^4 \left(\frac{5}{10} + \frac{9}{10}\right) = \frac{7}{5}\left(\frac{9}{10}\right)^4 = 0.9185$$

Example 3.30

Ten percent of screws produced in a certain factory turn out to be defective. Find the probability that in a sample of 10 screws chosen at random, exactly two will be defective.

Solution

$$p = \frac{1}{10},\ q = \frac{9}{10}$$

$$n = 10,\ r = 2$$

$$P(r) = {}^nC_r\ p^r\ q^{n-r}$$

$$P(2) = {}^{10}C_2 \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^{10-2}$$

$$= \frac{10 \times 9}{1 \times 2}\left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^8$$

$$= \frac{1}{2}\left(\frac{9}{10}\right)^9 = 0.1937$$

Example 3.31

The probability that a man aged 60 will live to be 70 is 0.65. What is the probability that out of 10 men, now 60, at least 7 will live to be 70?

Solution

The probability that a man aged 60 will live to be 70 = p = 0.65

$$q = 1 - p = 1 - 0.65 = 0.35$$

Number of men $n = 10$

Probability that at least 7 men will live to 70 = (7 or 8 or 9 or 10)

$$= P(7) + P(8) + P(9) + P(10)$$

$$= {}^{10}C_7\ q^3\ p^7 + {}^{10}C_8\ q^2\ p^8 + {}^{10}C_9\ q\ p^9 + p^{10}$$

$$= \frac{10 \times 9 \times 8}{1 \times 2 \times 3}(0.35)^3\ (0.65)^7 + \frac{10 \times 9}{1 \times 2}(0.35)^2\ (0.65)^8 + 10\ (0.35)\ (0.65)^9 + (0.65)^{10}$$

$$= (0.65)^7\ [120\ (0.35)^3 + 45\ (0.35)^2\ (0.65) + 10\ (0.35)\ (0.65)^2 + (0.65)^3]$$

$$= 0.5137$$

Mean and Standard Deviation of Binomial Distribution

We know that

$$(q+p)^n = q^n + n_{C_1}\, q^{n-1}\, p^1 + n_{C_2}\, q^{n-2}\, p^2 + n_{C_3}\, q^{n-3}\, p^3 + \ldots + n_{C_r}\, q^{n-r}\, p^r + \ldots + p^n$$

We know that mean is given by

$$\text{Mean} = \frac{\Sigma rf}{\Sigma f}$$

and $$\text{Variance} = \sigma^2 = \frac{\Sigma r^2 f}{\Sigma f} - \left(\frac{\Sigma fr}{\Sigma f}\right)^2$$

Consider the table

Successes *r*	**Frequency *f***	**Product *r f***	**Product $r^2 f$**
0	q^n	0	0
1	$n\, q^{n-1}\, p$	$n\, q^{n-1}\, p$	$n\, q^{n-1}\, p$
2	$\frac{n\,(n-1)}{2}\, q^{n-2}\, p^2$	$n\,(n-1)\, q^{n-2}\, p^2$	$2n\,(n-1)\, q^{n-2}\, p^2$
3	$\frac{n\,(n-1)\,(n-2)}{6}\, q^{n-3}\, p^3$	$\frac{n\,(n-1)\,(n-2)}{2}\, q^{n-3}\, p^3$	$\frac{3n\,(n-1)\,(n-2)}{2}\, q^{n-3}\, p^3$
. . .	. . .	. . .	. . .
n	p^n	$n\, p^n$	$n^2\, p^n$

Now,

$$\Sigma\, f\, r = n\, q^{n-1}\, p + n\,(n-1)\, q^{n-2}\, p^2 + \frac{n\,(n-1)\,(n-2)}{2}\, q^{n-3}\, p^3 + \ldots + n\, p^n$$

$$= n\, p\left[q^{n-1} + \frac{(n-1)}{1!}\, q^{n-2}\, p + \frac{(n-1)\,(n-2)}{2}\, q^{n-3}\, p^2 + \ldots + p^{n-1}\right]$$

$$= n\, p\,(q+p)^{n-1} \qquad \text{since } q+p=1$$

$$= n\, p$$

$$\Sigma\, f = q^n + n q^{n-1}\, p + \frac{n\,(n-1)}{2}\, q^{n-2}\, p^2 + \ldots + p^n$$

$$= (q+p)^n = 1 \qquad \text{since } q+p=1$$

$$\therefore \qquad \text{Mean} = \frac{\Sigma rf}{\Sigma f} = \frac{np}{1} = np$$

Also,

$$\Sigma f\, r^2 = 0 + n q^{n-1}\, p + 2n\,(n-1)\, q^{n-2}\, p^2 + \frac{3n\,(n-1)\,(n-2)}{2}\, q^{n-3}\, p^3 + \ldots + n^2\, p^n$$

$$= n\, p\left[q^{n-1} + \frac{2(n-1)}{1!}\, q^{n-2}\, p + \frac{3n\,(n-1)\,(n-2)}{2!}\, q^{n-3}\, p^2 + \ldots + n\, p^{n-1}\right]$$

$$= n\, p\left[q^{n-1} + \frac{(n-1)\, q^{n-2}\, p}{1!} + \frac{(n-1)\,(n-2)}{2!}\, q^{n-3}\, p^2 + \ldots + p^{n-1}\right.$$

$$\left. + \frac{(n-1)\, q^{n-2}\, p}{1!} + \frac{2(n-1)\,(n-2)}{2!}\, q^{n-3}\, p^2 + \ldots + (n-1)\, p^{n-1}\right]$$

$$= n\,p\left[q^{n-1} + (n-1)\,q^{n-2}\,p + \frac{(n-1)\,(n-2)}{2!}\,q^{n-3}\,p^2 + \ldots + p^{n-1}\right]$$

$$+ (n-1)\,n\,p^2\left[q^{n-2} + (n-2)\,q^{n-3}\,p + \frac{(n-2)\,(n-3)}{2!}\,q^{n-4}\,p^2 + \ldots + p^{n-2}\right]$$

$$= n\,p\,(q+p)^{n-1} + (n-1)\,n\,p^2\,(q+p)^{n-2}$$

$$= n\,p + (n-1)\,n\,p^2$$

$$= n\,p\,[1 + (n-1)\,p]$$

$$= n\,p\,[n\,p + (1-p)]$$

$$= n\,p\,[n\,p + q]$$

$$= n^2\,p^2 + n\,p\,q$$

Thus, Variance, $\sigma^2 = \dfrac{n^2\,p^2 + n\,p\,q}{1} - \left(\dfrac{n\,p}{1}\right)^2$

$$= n\,p\,q$$

$\therefore$ S. D., $\sigma = \sqrt{n\,p\,q}$

Hence for the binomial distribution,

Mean = $n\,p$

$$\mu_2 = \sigma^2 = n\,p\,q$$

Similarly, we can obtain the following moment generating functions.

$$\mu_3 = n\,p\,q\,(q-p)$$

$$\mu_4 = n\,p\,q\,[1 + 3\,p\,q\,(n-2)]$$

Hence the moment coefficient of skewness is

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

$$= \frac{[n\,p\,q\,(q-p)]^2}{(n\,p\,q)^3} = \frac{(q-p)^2}{n\,p\,q}$$

$$r_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{n\,p\,q}}$$

Coefficient of Kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{n\,p\,q\,[1 + 3\,p\,q\,(n-2)]}{(n\,p\,q)^2}$$

$$= 3 + \frac{1 - 6\,p\,q}{n\,p\,q}$$

$$\therefore \qquad r_2 = \beta_2 - 3 = \frac{1 - 6\,p\,q}{n\,p\,q}$$

Example 3.32

A die is tossed thrice. A success is getting 1 or 6 on a toss. Find the mean and variance of the number of successes.

Solution

$$n = 3,\ p = \frac{1}{3},\ q = \frac{2}{3}$$

Mean = $n\,p = 3 \times \frac{1}{3} = 1$

Variance = $n\,p\,q = 3 \times \frac{1}{3} \times \frac{2}{3} = \frac{2}{3}$

Recurrence Relation for the Binomial Distribution

By Binomial distribution

$$P(r) = {}^nC_r\ p^r\ q^{n-r} \qquad \ldots (3.39)$$

$$P(r+1) = {}^nC_{r+1}\ p^{r+1}\ q^{n-r-1} \qquad \ldots (3.40)$$

On dividing Eq. (3.40) by Eq. (3.39), we get

$$\frac{P(r+1)}{P(r)} = \frac{{}^nC_{r+1}}{{}^nC_r}\ \frac{p^{r+1}\ q^{n-r-1}}{p^r\ q^{n-r}}$$

$$= \frac{n\,(n-1)\,(n-2)\ldots(n-r)}{(r+1)!}\ \frac{r!}{n\,(n-1)\,(n-2)\ldots(n-r+1)}\ \frac{p}{q}$$

$$= \frac{(n-r)}{(r+1)}\ \frac{p}{q}$$

$$P(r+1) = \frac{(n-r)}{(r+1)}\ \frac{p}{q}\ P(r),$$

which is the recurrence relation for the Binomial Distribution.

SAQ 7

(a) If 20% of the bolts produced by a machine are defective, determine the probability that out of 4 bolts chosen at random

(i) 1

(ii) 0

(iii) At most 2

bolts will be defective.

(b) The probability density function of a variable X is

X	0	1	2	3	4	5	6
$P(X)$	k	$3k$	$5k$	$7k$	$9k$	$11k$	$13k$

(i) Find $P(X < 4)$, $P(X \geq 5)$, $P(3 < X \leq 6)$.

(ii) What will be the minimum value of k so that $P(X \leq 2) > 3$?

(c) A random variable x has the following probability function :

Values of x	– 2	– 1	0	1	2	3
$P(x)$	0.1	k	0.2	$2k$	0.3	k

Find the value of k and calculate mean and variance.

(d) (i) If a new drug is found to be effective 40% of the time, then what is the probability that in a random sample of 4 patients, it will be effective on 2 of them?

(ii) In a manufacturing process, a packaging machine produces 5% defective packages. Find the mean and the standard deviation of the number of defective packages in a random sample of 60 packages.

(e) A multiple choice test contains 6 questions. Each question has 3 answers of which only 1 is the correct answer. The student has no idea as to which of the alternatives is the correct one. The student rolls a fair dice. If face 1 or 2 show up, he selects answer (a). If face 3 or 4 show up, he selects answer (b) and if face 5 or 6 show up, he selects answer (c). Find the probability that he will get

(i) exactly 4 correct answers,

(ii) no correct answer

(iii) at most 2 correct answers.

3.12.2 The Poisson Distribution

Poisson distribution is a limiting case of the Binomial distribution when p (or q) is very small and n is large enough.

In Binomial distribution

$$P(r) = {}^nC_r\, q^{n-r}\, p^r \qquad \text{(mean } m = n\,p\text{)}$$

$$= {}^nC_r\, (1-p)^{n-r}\, p^r \qquad \text{(since } m = n\,p,\ \ p = \frac{m}{n}\text{)}$$

$$= {}^nC_r \left(1 - \frac{m}{n}\right)^{n-r} \left(\frac{m}{n}\right)^r \qquad (m \text{ is constant})$$

$$= \frac{n\,(n-1)\,(n-2)\ldots(n-\overline{r-1})}{r!} \left(\frac{m}{n}\right)^r \left(1 - \frac{m}{n}\right)^{n-r}$$

$$= \frac{\frac{n}{n}\left(\frac{n}{n} - \frac{1}{n}\right)\left(\frac{n}{n} - \frac{2}{n}\right)\ldots\left(\frac{n}{n} - \frac{r-1}{n}\right) m^r \left(1 - \frac{m}{n}\right)^n}{r! \qquad \left(1 - \frac{m}{n}\right)^r}$$

$$= \frac{1\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\ldots\left(1 - \frac{r-1}{n}\right) m^r \left(1 - \frac{m}{n}\right)^n}{r! \qquad \left(1 - \frac{m}{n}\right)^r}$$

Taking limits when n tends to infinity

$$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = \lim_{n\to\infty}\left[\left(1-\frac{m}{n}\right)^{\frac{-n}{m}}\right]^{-m} = e^{-m}$$

$$P(r) = \frac{m^r}{r!}e^{-m}$$

$$P(r) = \frac{e^{-m}\,.\,m^r}{r!} \qquad \ldots (3.41)$$

which is the Poisson probability mass function.

The Poisson probability distribution is often used as a model of the number of arrivals at a facility within a given period of time. For instance, a random variable might be defined as the number of telephone calls coming into an airline reservation system during a period of 15 minutes.

If the mean number of arrivals during a 15-minute interval is known, the Poisson probability mass function given by Eq. (3.41) can be used to compute the probability of r arrivals.

Mean of Poisson Distribution

We know that

$$P(r) = \frac{e^{-m}\,m^r}{r!}$$

Successes r	**Frequency** f	**Product** rf
0	$\frac{e^{-m}\,m^0}{0!}$	0
1	$e^{-m}\,m^1$	$e^{-m}\,.\,m$
2	$\frac{e^{-m}\,m^2}{2!}$	$e^{-m}\,.\,m^2$
3	$\frac{e^{-m}\,m^3}{3!}$	$\frac{e^{-m}\,.\,m^3}{2!}$
. . . . .	. . . . .	. . . . .
r	$\frac{e^{-m}\,m^r}{r!}$	$\frac{e^{-m}\,.\,m^r}{(r-1)!}$
. . . . .	. . . . .	. . . . .

$$\Sigma f r = 0 + e^{-m}\,.\,m + e^{-m}\,.\,m^2 + e^{-m}\,.\,\frac{m^3}{2!} + \ldots + e^{-m}\,.\,\frac{m^r}{(r-1)!} + \ldots$$

$$= e^{-m}\,.\,m\left[1 + \frac{m}{1!} + \frac{m^2}{2!} + \ldots + \frac{m^{r-1}}{(r-1)!} + \ldots\right]$$

$$= m\,.\,e^{-m}\,.\,[e^m]$$

$$= m$$

Also, $\Sigma f = 1$

$$\therefore \qquad \text{Mean} = \frac{\Sigma f r}{\Sigma f} = \frac{m}{1}$$

Second Method

In Binomial distribution,

Mean = $n\,p$

Poisson mean = $\lim\limits_{n\to\infty} n\,p = \lim\limits_{n\to\infty} n\,.\,\dfrac{m}{n} = \lim\limits_{n\to\infty} m = m$

Example 3.33

Fit a Poisson distribution to the following :

x	0	1	2	3	4
f	192	100	24	3	1

Solution

x	f	fx
0	192	0
1	100	100
2	24	48
3	3	9
4	1	4
Total	320	161

$$\text{Mean} = \frac{\sum f\,x}{\sum f}$$

$$= \frac{161}{320} = 0.5 \qquad \text{nearly}$$

$$P(r) = \frac{e^{-m}\,.\,m^r}{r!}$$

$$= \frac{e^{-0.5}\,.\,(0.5)^r}{r!}$$

This is the required Poisson distribution.

Standard Deviation of Poisson Distribution

$$P(r) = \frac{e^{-m}\,m^r}{r!}$$

$$\sum f = 1,\ \sum f\,r = m$$

Successes r	**Frequency** f	**Product** $r\,f$	**Product** $r^2 f$
0	$\dfrac{e^{-m}\,m^0}{0!}$	0	0
1	$\dfrac{e^{-m}\,m^1}{1!}$	$e^{-m}\,.\,m$	$e^{-m}\,.\,m$

2	$\frac{e^{-m} m^2}{2!}$	$e^{-m} . m^2$	$2e^{-m} . m^2$
3	$\frac{e^{-m} m^3}{3!}$	$\frac{e^{-m} . m^3}{2!}$	$\frac{3e^{-m} . m^3}{2!}$
. . . . .	. . . . .	. . . . .	. . . . .
r	$\frac{e^{-m} m^r}{r!}$	$\frac{e^{-m} . m^r}{(r-1)!}$	$\frac{r\, e^{-m} . m^r}{(r-1)!}$
. . . . .	. . . . .	. . . . .	. . . . .

$$\sum f r^2 = 0 + e^{-m} . m + 2e^{-m} . m^2 + 3 . e^{-m} . \frac{m^3}{2!} + \ldots + \frac{re^{-m} . m^r}{(r-1)!} + \ldots$$

$$= m . e^{-m} \left[1 + 2m + \frac{3m^2}{2!} + \frac{4m^3}{3!} + \ldots + \frac{r . m^{r-1}}{(r-1)!} + \ldots \right]$$

$$= m . e^{-m} \left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \ldots + \frac{m^{r-1}}{(r-1)!} + \ldots \right.$$

$$\left. m + \frac{2m^2}{2!} + \frac{3m^3}{3!} + \ldots + \frac{(r-1)\, m^{r-1}}{(r-1)!} + \ldots \right]$$

$$= m . e^{-m} \left[\left\{ 1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \ldots + \frac{m^{r-1}}{(r-1)!} + \ldots \right\} \right.$$

$$\left. + m \left\{ 1 + \frac{m}{1!} + \frac{m^2}{2!} + \ldots + \frac{m^{r-2}}{(r-2)!} + \ldots \right\} \right]$$

$$= m . e^{-m} [e^m + m\, e^m]$$

$$= m + m^2$$

$$\sigma^2 = \frac{\sum f r^2}{\sum f} - \left(\frac{\sum f r}{\sum f} \right)^2 = \frac{m + m^2}{1} - (m)^2 = m$$

$$\sigma = \sqrt{m}$$

Second Method

$\sigma^2 = n\, p\, q$ (In Binomial distribution)

$\sigma^2 = \lim_{n \to \infty} n\, p\, q$ (In Poisson distribution)

$$= \lim_{n \to \infty} n \left(\frac{m}{n} \right) \left(1 - \frac{m}{n} \right)$$

$$= \lim_{n \to \infty} m \left(1 - \frac{m}{n} \right)$$

$$= m$$

Hence mean and variance of a Poisson distribution are each equal to m. Similarly, we can obtain

$$\mu_3 = m, \quad \mu_4 = 3m^2 + m$$

$$\beta_1 = \frac{1}{m}, \quad \beta_2 = 3 + \frac{1}{m}$$

$$\gamma_1 = \frac{1}{\sqrt{m}}, \quad \gamma_2 = \frac{1}{m}$$

Example 3.34

If the variance of the Poisson distribution is 2, find the probabilities for $r = 1, 2, 3, 4$ from the recurrence relation of the Poisson distribution. Also find $P(r \geq 4)$.

Solution

Variance = $m = 2$

$$P(r) = \frac{e^{-m} \cdot m^r}{r!}$$

$$P(r) = \frac{e^{-2} \cdot 2^r}{r!}$$

$$P(1) = \frac{e^{-2} \cdot 2^1}{1!} = 0.271$$

$$P(2) = \frac{e^{-2} \cdot 2^2}{2!} = 0.271$$

$$P(3) = \frac{e^{-2} \cdot 2^3}{3!} = 0.180$$

$$P(4) = \frac{e^{-2} \cdot 2^4}{4!} = 0.090$$

$$\begin{aligned} P(r \geq 4) &= 1 - [P(0) + P(1) + P(2) + P(3)] \\ &= 1 - [0.135 + 0.271 + 0.271 + 0.180] \\ &= 1 - [0.857] \\ &= 0.143. \end{aligned}$$

SAQ 8

(a) Assume that the probability of an individual coal miner being killed in a mine accident during a year is 1/2400. Use appropriate distribution to calculate the probability that in mine employing 200 miners, there will be at least one such accident in a year.

(b) In a certain factory producing cycle tyres there is a small chance of 1 in 500 tyres to be defective. The tyres are supplied in lots of 10. Using Poisson distribution, calculate the approximate number of lots containing no defective, one defective, and two defective tyres, respectively, in a consignment of 10,000 lots.

(c) Customers arrive at a photocopying machine at an average rate of two every 10 minutes. The number of arrivals is distributed according to a Poisson distribution. What is the probability that there will be :

(g) no arrivals during any period of ten minutes?

(iii) exactly one arrival during this time period?

(iv) more than two arrivals during this time period?

3.12.3 The Normal Distribution

So far we have dealt with discrete distributions where the variate takes only the integral values. But the variates like temperature, heights and weights can take all values in a given interval. Such variables are called continuous variables.

Let $f(x)$ be a continuous function.

Here $f(x)$ is called probability density function if

(i) $f(x) \geq 0$ for every value of x

(ii) $\int_{-\infty}^{\infty} f(x)\, dx = 1$

(iii) $\int_{a}^{b} f(x)\, dx = P(a < x < b)$

For continuous probability density function $f(x)$,

$$\text{Mean} = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

and Variance (from mean $\bar{x}$) $= \int_{-\infty}^{\infty} (x - \bar{x})^2\, f(x)\, dx$

The most widely used continuous probability distribution in statistics is the normal probability distribution.

The normal distribution is the most useful and popular probability function. It can be completely specified by two parameters : the mean (μ) and the variance (σ^2). The probability density function for normal distribution is given as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \qquad \ldots (3.42)$$

The mean of a normal distribution locates the center of the function and can be any real number. The variance of a normal distribution measures the variability and can be any positive real number. The normal distribution is symmetrical about the mean as shown in Figure 11.1. The area under the curve between any two points represents the probability that the random variable will lie between these two points. Obviously, the total area under the curve is 1. The normal curve has the property that 68% of the population lie within $\pm\,\sigma$ limits, 95.4% within $\pm\,2\sigma$ limits and 99.7% within $\pm\,3\sigma$ limits.

Figure 3.13 : Area Under Normal Curve

It is convenient to look for the probability values from standard table. Table 3.14 is the standard normal table. To use a table we first convert the normal distribution to a standard normal distribution. A standard normal distribution is one that has a mean of 0

and a standard deviation of 1. Any normal distribution can be transformed to a standard normal distribution using the following transformation :

$$z = \frac{x - \mu}{\sigma} \qquad \ldots (3.43)$$

The probability may be found from Table 3.14. Let us first learn to read this table. Suppose you want to find out the area under the normal curve to the left of z =1.54. First go to the row corresponding to z = 1.5 and in that row move to the column corresponding to 0.04. Value in that particular cell is 0.9382. Hence, the required area is 0.9382.

An Example Illustrating the Use of Table 3.14

For a probability distribution, $\mu = 100$ and $\sigma = 15$. It is required to find out the probabilities $P(x < 125)$ and $P(x < 75)$. The probabilities are calculated as follows :

(i) For random variable X less than 125, $Z = (125 - 100)/15 = 1.66$.

From the tables, $P(X < 125) = P(Z = 1.66) = 0.9515$. This represents the percentage area (shaded) under the normal curve as shown in Figure 3.14(a).

(ii) The random variable X is less than 75, i.e. $P(X < 75)$

$Z = (75 - 100)/15 = -1.66$. The standard tables provide values for positive Z only. Since we notice that the areas about mean are symmetric, the probability $P(x < 75)$ is equivalent to $P(x > 125)$. From Table 3.14, $P(x < 125) = 0.9515$. Therefore,

$$P(x < 75) = P(x > 125) = 1 - P(x < 125) = 1 - 0.9515 = 0.0485.$$

This represents percentage area (shaded) under normal curve as shown in Figure 3.14(b).

(a) Area Under Curve When $X < 125$ **(b) Area Under Curve When $X < 75$**

Figure 3.14

Example 3.35

A function $f(x)$ is defined as follows :

$$f(x) = \begin{cases} 0 & x < 2 \\ \frac{1}{18}(2x+3) & 2 \le x \le 4 \\ 0 & x > 4 \end{cases}$$

Show that it is probability density function.

Solution

$$f(x) = \begin{cases} 0 & x < 2 \\ \frac{1}{18}(2x+3) & 2 \le x \le 4 \\ 0 & x > 4 \end{cases}$$

If $f(x)$ is a probability density function then $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

Here $\int_2^4 \frac{1}{18}(2x+3)\, dx = \frac{1}{18}(x^2 + 3x) \Big|_2^4$

$$= \frac{1}{18}[(16+12) - (4+6)]$$

$$= \frac{1}{18}[28 - 10) = \frac{18}{18} = 1$$

Hence the given function is a probability density function.

Table 3.14 : Area Under the Normal Curve

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.98
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.98
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.98
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.99
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.99
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.99
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.99
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.99
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.99
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.99
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.99

Example 3.36

The diameter of an electric cable is assumed to be continuous random variate with p. d. f. (Probability density function).

$$f(x) = 6x\,(1 - x), \qquad 0 \le x \le 1$$

(i) Verify that above is probability density function.

(ii) Find the mean and variance.

Solution

(i) $$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 6x\,(1 - x)\,dx$$

$$= \int_0^1 (6x - 6x^2)\,dx = \left(\frac{6x^2}{2} - \frac{6x^3}{3}\right)_0^1$$

$$= \frac{6}{2} - \frac{6}{3} = 3 - 2 = 1$$

Hence the given function is a probability function.

(ii) Mean $= \int_{-\infty}^{\infty} x\,f(x)\,dx$

$$= \int_0^1 x\,.\,6x\,(1 - x)\,dx = 6\int_0^1 (x^2 - x^3)\,dx$$

$$= 6\left(\frac{x^3}{3} - \frac{x^4}{4}\right)_0^1 = 6\left(\frac{1}{3} - \frac{1}{4}\right) = 6\left(\frac{1}{12}\right) = \frac{1}{2}$$

Variance $= \int_{-\infty}^{\infty} (x - \bar{x})^2\,.\,f(x)\,dx$

$$= \int_0^1 \left(x - \frac{1}{2}\right)^2 6x\,(1 - x)\,dx = \int_0^1 \left(x^2 + \frac{1}{4} - x\right)(6x - 6x^2)\,dx$$

$$= \int_0^1 \left(6x^3 - 6x^4 + \frac{6x}{4} - \frac{6x^2}{4} - 6x^2 + 6x^3\right) dx$$

$$= \left(\frac{6x^4}{4} - \frac{6x^5}{5} + \frac{6x^2}{8} - \frac{6x^3}{12} - \frac{6x^3}{3} + \frac{6x^4}{4}\right)_0^1$$

$$= \frac{6}{4} - \frac{6}{5} + \frac{6}{8} - \frac{6}{12} - \frac{6}{3} + \frac{6}{4}$$

$$= \frac{1}{20}$$

Mean for Normal Distribution

$$\text{Mean} = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} . x\, dx$$

Putting $\frac{x}{\sigma} = t$, we get

$$\text{Mean} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} (t\,\sigma)\,(\sigma\, dt)$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t\, e^{-\frac{t^2}{2}}\, dt$$

$$= \frac{\sigma}{\sqrt{2\pi}} \left[e^{-\frac{t^2}{2}} \right]_{-\infty}^{+\infty}$$

$$= \frac{\sigma}{\sqrt{2\pi}} [0]$$

$$= 0.$$

Example 3.37

Fit a normal curve to the following table :

Length of Line (cm)	8.60	8.59	8.58	8.57	8.56	8.55	8.54	8.53	8.52
Frequency	2	3	4	9	10	8	4	1	1

Solution

Let $a = 8.56$

x	f	$d = (x - a)$	fd	fd^2
8.60	2	0.04	0.08	0.0032
8.59	3	0.03	0.09	0.0027
8.58	4	0.02	0.08	0.0016
8.57	9	0.01	0.09	0.0009
8.56	10	0	0	0
8.55	8	– 0.01	– 0.08	0.0008
8.54	4	– 0.02	– 0.08	0.0016
8.53	1	– 0.03	– 0.03	0.0009
8.52	1	– 0.04	– 0.04	0.0016
	$\Sigma f = 42$		$\Sigma fd = 0.11$	$\Sigma fd^2 = 0.0133$

$$\text{Mean } (\mu) = a + \frac{\Sigma fd}{\Sigma f} = 8.56 + \frac{0.11}{42}$$

$$= 8.56262$$

$$\text{S. D. } (\sigma) = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$$

$$(\sigma) = \sqrt{\frac{0.0133}{42} - \left(\frac{0.11}{42}\right)^2}$$

$$= 0.0176$$

Hence the equation of the normal curve fitted to the given data is,

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (-\propto \le x \le \propto)$$

where $\mu = 8.56262, \ \sigma = 0.0176$

SAQ 9

(a) 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 more than 60 marks?

(b) The heights of soldiers are normally distributed. If 11.51% of the soldiers are taller than 70.4 inches and 9.68% are shorter than 65.4 inches, find the mean and the standard deviation for the data of heights of soldiers.

3.13 STATISTICAL INFERENCE AND SAMPLE STATISTICS

There are many other discrete and continuous probability distributions. Other widely used discrete distributions include the geometric, the hyper-geometric, and the negative binomial; other commonly used continuous distributions include the uniform, exponential, gamma, chi-square, beta, *t*, and *F*.

It is often of interest to learn about the characteristics of a large group of elements such as individuals, households, buildings, products, parts, customers, and so on. All the elements of interest in a particular study form the **population**. Because of time, cost, and other considerations, data often cannot be collected from every element of the population. In such cases, a subset of the population, called a **sample**, is used to provide the data. Data from the sample are then used to develop estimates of the characteristics of the (larger) population. The process of using a sample to make inferences about a population is called **statistical inference**.

Characteristics such as the population mean, the population variance, and the population proportion are called **parameters** of the population. Characteristics of the sample such as the sample mean, the **sample variance**, and the sample proportion are called sample statistics. There are two types of estimates: point and interval. A *point estimate* is a value of a sample statistic that is used as a single estimate of a population parameter. No statements are made about the quality or precision of a point estimate. Statisticians prefer interval estimates because interval estimates are accompanied by a statement concerning the degree of confidence that the interval contains the population parameter being estimated. Interval estimates of population parameters are called *confidence intervals*.

3.13.1 Statistical Quality Control

Statistical quality control refers to the use of statistical methods in the monitoring and maintaining of the quality of **products** and **services**. One method, referred to as acceptance sampling, can be used when a decision must be made to accept or reject a group of parts or items based on the quality found in a sample. A second method,

referred to as **statistical process control**, uses graphical displays known as control charts to determine whether a **process** should be continued or should be adjusted to achieve the desired quality.

3.13.2 Acceptance Sampling

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

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

the two control limits. Standard practice is to set the control limits at three standard deviations above and below the process mean. The process can be sampled periodically. As each sample is selected, the value of the sample mean is plotted on the control chart. If the value of a sample mean is within the control limits, the process can be continued under the assumption that the quality standards are being maintained. If the value of the sample mean is outside the control limits, an out-of-control conclusion points to the need for corrective action in order to return the process to acceptable quality levels.

The field of sample survey methods is concerned with effective ways of obtaining sample data. The three most common types of sample surveys are mail surveys, telephone surveys, and personal interview surveys. All of these involve the use of a questionnaire, for which a large body of knowledge exists concerning the phrasing, sequencing, and grouping of questions. There are other types of sample surveys that do not involve a questionnaire. For example, the sampling of accounting records for audits and the use of a computer to sample a large database are sample surveys that use direct observation of the sampled units to collect the data.

A goal in the design of sample surveys is to obtain a sample that is representative of the population so that precise inferences can be made. Sampling error is the difference between a population parameter and a sample statistic used to estimate it. For example, the difference between a population mean and a sample mean is sampling error. Sampling error occurs because a portion, and not the entire population, is surveyed. Probability sampling methods, where the probability of each unit appearing in the sample is known, enable statisticians to make probability statements about the size of the sampling error. Nonprobability sampling methods, which are based on convenience or judgment rather than on probability, are frequently used for cost and time advantages. However, one should be extremely careful in making inferences from a non-probability sample; whether or not the sample is representative is dependent on the judgment of the individuals designing and conducting the survey and not on sound statistical principles. In addition, there is no objective basis for establishing bounds on the sampling error when a non-probability sample has been used.

Most governmental and professional polling surveys employ probability sampling. It can generally be assumed that any survey that reports a plus or minus margin of error has been conducted using probability sampling. Statisticians prefer probability sampling methods and recommend that they be used whenever possible. A variety of probability sampling methods are available. A few of the more common ones are reviewed here.

Stratified simple random sampling is a variation of simple random sampling in which the population is partitioned into relatively homogeneous groups called strata and a simple random sample is selected from each stratum. The results from the strata are then aggregated to make inferences about the population. A side benefit of this method is that inferences about the subpopulation represented by each stratum can also be made.

Cluster sampling involves partitioning the population into separate groups called clusters. Unlike in the case of stratified simple random sampling, it is desirable for the clusters to be composed of heterogeneous units. In single-stage cluster sampling, a simple random sample of clusters is selected, and data are collected from every unit in the sampled clusters. In two-stage cluster sampling, a simple random sample of clusters is selected

and then a simple random sample is selected from the units in each sampled cluster. One of the primary applications of cluster sampling is called area sampling, where the clusters are counties, townships, city blocks, or other well-defined geographic sections of the population.

3.14 CORRELATION

Types of Correlation

Positive Correlation

Increase in the value of one variable results in a corresponding increase in the value of the other variable. This is also known as direct correlation.

Example : Amount of rainfall and yield of crop (up to a certain extent); family income and expenditure on luxury items.

Negative Correlation

Increase (decrease) in the value of one variable results on an average decrease(increase) in the value of the other variable. This is also known as indirect correlation.

Example : Volume and pressure of a perfect gas.

Linear Correlation

Corresponding to a unit change in one variable there is a constant change in the other variable over the entire range of the values.

X :	1	2	3	4	5
Y :	3	6	9	12	15

thus $y = 3x$

Non-Linear Corrrelation

Corresponding to a unit change in one variable, the other variable does not change at a constant rate but at a fluctuating rate.

Methods to Determine Correlation

(i) Scatter Diagram

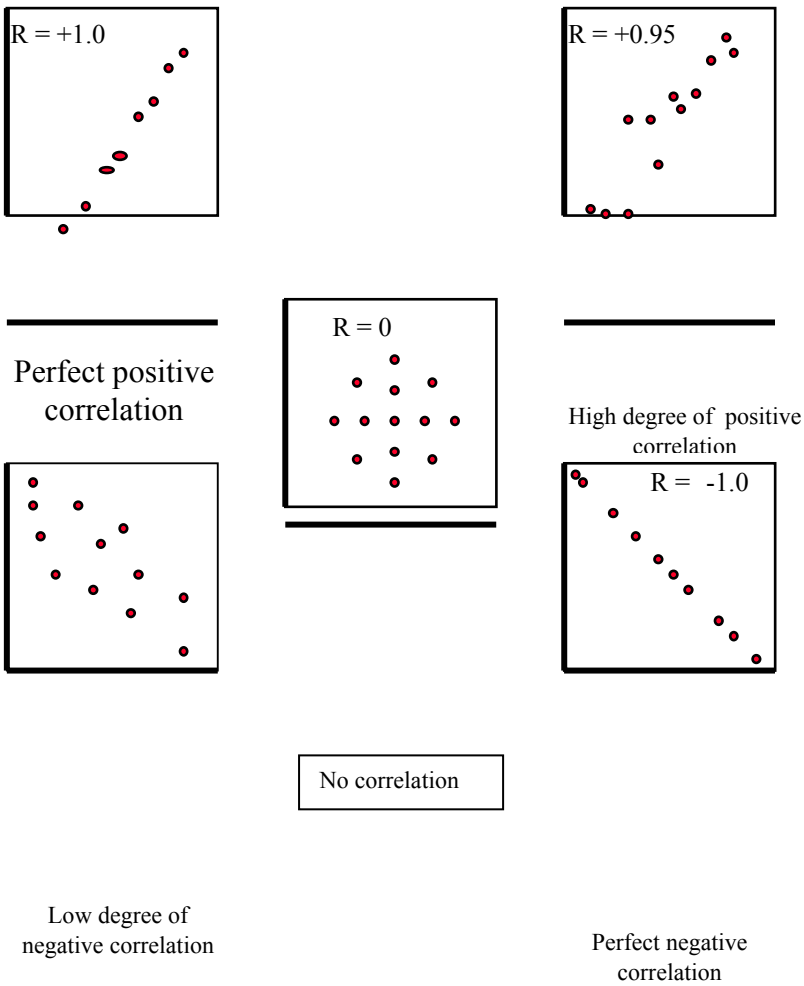
(ii) Karl Pearson's coefficient of correlation

(iii) Rank Method

Scatter Diagram

It helps us to determine the correlation between two variables by means of a diagram. The method is to plot the dependent variable on the y-axis and the independent variable on the x-axis. The following patterns may be observed after plotting the points.

Scatter Diagram



Karl Pearson's Coefficient of Correlation

A correlation coefficient is a number between (– 1) and (1) which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1; if there is positive correlation, whenever one variable has a high (low) value, so does the other. If there is a perfect linear relationship with negative slope between the two variables, we have a correlation coefficient of (– 1); if there is negative correlation, whenever one variable has a high (low) value, the other has a low (high) value. A correlation coefficient of 0 means that there is no linear relationship between the variables.

Karl Pearson's product moment correlation coefficient, usually denoted by r, is one example of a correlation coefficient.

$$r = \frac{\text{Cov}\,(x, y)}{\sigma_x\ \sigma_y} \qquad \ldots (3.42)$$

where r = coefficient of correlation

Cov (x, y) = covariance between x and y

σ_x and σ_y = standard deviations of x and y

$$\text{Cov}\,(x, y) = \frac{1}{n} \sum (x - \bar{x})\,(y - \bar{y})$$

$$\sigma_x = \sqrt{\left(\frac{1}{n}\right) \sum (x - \bar{x})^2}$$

$$\sigma_y = \sqrt{\left(\frac{1}{n}\right) \Sigma (y - \bar{y})^2}$$

Thus, $$r = \frac{\left(\frac{1}{n}\right) \Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\left(\frac{1}{n}\right) \Sigma (x - \bar{x})^2 \cdot \left(\frac{1}{n}\right) \Sigma (y - \bar{y})^2}}$$

Thus, $$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma (x - \bar{x})^2 \, \Sigma (y - \bar{y})^2}}$$

Let $dx = x - \bar{x}$, $dy = (y - \bar{y})$

Thus, $$r = \frac{\Sigma\, dx\, dy}{\sqrt{\Sigma dx^2 \, \Sigma dy^2}}$$

Consider, $$\text{Cov}(x, y) = \frac{1}{n} \Sigma (x - \bar{x})(y - \bar{y})$$

$$= \left(\frac{1}{n}\right) \Sigma (xy - \bar{x}y - x\bar{y} + \bar{x}\,\bar{y})$$

$$= \left(\frac{1}{n}\right) \Sigma\, xy - \bar{x}\left(\frac{1}{n}\right) \Sigma\, y - \bar{y}\left(\frac{1}{n}\right) \Sigma\, x + \left(\frac{1}{n}\right) \Sigma\, \bar{x}\,\bar{y}$$

$$= \left(\frac{1}{n}\right) \Sigma\, xy - \bar{x}\,\bar{y} - \bar{y}\,\bar{x} + \frac{1}{n}\, n\, \bar{x}\,\bar{y}$$

$$= \left(\frac{1}{n}\right) \Sigma\, xy - \bar{x}\,\bar{y} - \bar{x}\,\bar{y} + \bar{x}\,\bar{y}$$

$$= \left(\frac{1}{n}\right) \Sigma\, xy - \left(\frac{\Sigma x}{n}\right)\left(\frac{\Sigma y}{n}\right)$$

$$= \left(\frac{1}{n^2}\right) [n \Sigma\, xy - (\Sigma\, x)(\Sigma\, y)] \qquad \ldots (3.43)$$

$$\sigma_x = \left(\frac{1}{n}\right) \Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2 \qquad \ldots (3.44)$$

$$\sigma_y = \left(\frac{1}{n}\right) \Sigma y^2 - \left(\frac{\Sigma y}{n}\right)^2 \qquad \ldots (3.45)$$

Substituting Eqs. (3.43) and (3.44) in Eq. (3.42), we get

$$r = \frac{n \Sigma\, xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma (x)^2 - \Sigma (x)^2 \,.\, n \Sigma (y)^2 - \Sigma (y)^2}} \qquad \ldots (3.46)$$

Example 3.38

Calculate the Karl Pearson's coefficient of correlation from the following data

x :	28	45	40	38	35	33	40	32	36	33
y :	23	34	33	34	30	26	28	31	36	35

Solution

x	x^2	y	y^2	xy

28	784	23	529	644
45	2025	34	1156	1530
40	1600	33	1089	1320
38	1444	34	1156	1292
35	1225	30	900	1050
33	1089	26	676	858
40	1600	28	784	1120
32	1024	31	961	992
36	1296	36	1296	1296
33	1089	35	1225	1155
Σ360	Σ13176	Σ310	Σ9772	Σ11257

Now the formula for r can also be written as :

$$r = \frac{\left(\Sigma \frac{xy}{n}\right) - \left(\Sigma \frac{x}{n}\right)\left(\Sigma \frac{y}{n}\right)}{\sqrt{\Sigma\left(\frac{x^2}{n}\right) - \Sigma\left(\frac{x}{n}\right)^2 - \Sigma\left(\frac{y^2}{n}\right) - \Sigma\left(\frac{y}{n}\right)^2}}$$

$$= \frac{9.7}{(4.65)\ (4.02)} = 0.5189\,.$$

Spearman's Rank Correlation Coefficient

The Spearman rank correlation coefficient is one example of a correlation coefficient. It is usually calculated on occasions when it is not convenient, economic, or even possible to give actual values to variables, but only to assign a rank order to instances of each variable.

In calculating this coefficient, we use the Greek letter 'rho' or ρ
The formula used to calculate this coefficient is :

$$\rho = 1 - \frac{6\Sigma\, d^2}{n\,(n^2 - 1)} \qquad \ldots (3.47)$$

where 'ρ' denotes rank coefficient of correlation and '*d*' refers to the difference of ranks between paired items in two series.

Example 3.39

Researchers at the European Centre for Road Safety Testing are trying to find out how the age of cars affects their braking capability. They test a group of ten cars of differing ages and find out the minimum stopping distances that the cars can achieve. The results are set out in the table below :

Table 3.10 : Car Ages and Stopping Distances

Car	Age (months)	Minimum Stopping at 40 kph (metres)

A	9	28.4
B	15	29.3
C	24	37.6
D	30	36.2
E	38	36.5
F	46	35.3
G	53	36.2
H	60	44.1
I	64	44.8
J	76	47.2

Solution

To process this information we must, firstly, place the ten pieces of data into order, or rank them according to their age and ability to stop. It is then possible to process these ranks.

Table 3.11 : Ranked Data from Table 3.10 Above

Car	Age (months)	Minimum Stopping at 40 kph (metres)	Age rank	Stopping rank
A	9	28.4	1	1
B	15	29.3	2	2
C	24	37.6	3	7
D	30	36.2	4	4.5
E	38	36.5	5	6
F	46	35.3	6	3
G	53	36.2	7	4.5
H	60	44.1	8	8
I	64	44.8	9	9
J	76	47.2	10	10

Notice that the ranking is done here in such a way that the youngest car and the best stopping performance are rated top and vice versa. There is no strict rule here other than the need to be consistent in your rankings. Notice also that there were two values the same (for cars D and G) in terms of the stopping performance of the cars tested. They occupy 'tied ranks' and must share, in this case, ranks 4 and 5. This means they are each ranked as 4.5, which is the mean average of the two ranking places. It is important to remember that this works despite the number of items sharing tied ranks. For instance, if five items shared ranks 5, 6, 7, 8 and 9, then they would each be ranked 7 - the mean of the tied ranks.

Now we can start to process these ranks to produce the following table :

Table 3.12 : Analysis of Data from Table 3.11

Car	Age (mths)	Stopping distance	Age rank	Stopping rank	*D*	d^2
A	9	28.4	1	1	0	0
B	15	29.3	2	2	0	0
C	24	37.6	3	7	4	16
D	30	36.2	4	4.5	0.5	0.25
E	38	36.5	5	6	1	1
F	46	35.3	6	3	-3	9
G	53	36.2	7	4.5	-2.5	6.25
H	60	44.1	8	8	0	0
I	64	44.8	9	9	0	0
J	76	47.2	10	10	0	0
						$\Sigma d^2 = 32.5$

Note that the two extra columns introduced into the new table are Column 6, '*d*', the difference between stopping distance rank and age rank; and Column 7, 'd^2', Column 6 entries squared. These squared figures are summed at the foot of Column 7.

Calculation of Spearman Rank Correlation Coefficient (*r*) is :

$$r = 1 - \frac{6\sum d^2}{n\,(n^2 - 1)}$$

Number in sample (*n*) = 10

$$r = 1 - \frac{6 \times 32.5}{10(10^2 - 1)}$$

$$r = 1 - \left(\frac{195}{10 \times 99}\right)$$

$$r = 1 - 0.197$$

$$r = 0.803$$

Interpretation of the Value of the Correlation Coefficient

(a) The correlation coefficient is always between –1 and + 1.

(b) The closer the correlation is to + 1 or – 1, the closer to a perfect linear relationship.

(c) – 1.0 to – 0.7 high degree of negative correlation.

(d) – 0.7 to – 0.3 low degree of negative correlation.

(e) – 0.3 to + 0.3 little or no correlation.

(f) + 0.3 to + 0.7 low degree of positive correlation.

(g) + 0.7 to + 1.0 high degree of positive correlation.

SAQ 10

(a) The data given represents the heights and weights of 10 girls.

Height : (in inches)	62	78	65	55	63	70	68	62	60	75
Weight (in kgs.) :	50	65	63	50	53	60	59	51	55	65

Draw a scatter diagram and indicate whether the correlation is positive or negative.

(b) The following table gives the number of motor vehicles with licenses and the number of motor vehicle accidents in a city. Calculate the coefficient of correlation between the number of motor vehicles and the number of motor vehicle accidents.

No. of vehicles in '000s :	2.6	2.8	2.9	3.1	3.2	2.3	2.5	1.8
No. of accidents in '00s :	5.9	6.0	6.2	6.2	7.6	7.0	7.5	5.5

(c) Calculate the coefficient of correlation between the export of raw material and import of finished goods from the following data

Export of Raw Material (in crores of Rs.)	42	44	58	55	89	98	66
Import of Finished Goods (in crores of Rs.)	56	59	53	58	65	78	58

(d) Nine girls obtained the following marks in the preliminary examination (x) and in the final examination (y). Calculate the correlation coefficient between x and y.

x	51	53	73	46	50	60	47	36	60
y	49	72	74	44	58	66	50	30	55

(e) Find the correlation coefficient between the demand and price of a certain commodity

Demand	65	66	67	67	68	69	70	72
Price	67	68	65	68	72	72	69	71

(f) The rank of 10 students in two subjects A and B are as follows :

A	3	5	8	4	7	10	2	1	6	9
B	6	4	9	8	1	2	3	10	5	7

What is the coefficient of rank correlation?

(g) Two judges in a beauty competition rank the 12 entries as follows :

X	1	2	3	4	5	6	7	8	9	10	11
Y	12	9	6	10	3	5	4	7	8	2	11

What degree of agreement is there between the two judges.

3.15 REGRESSION

The term regression was first used by Galton. He found that although tall fathers have tall sons and short fathers have short sons, the average height of sons of tall fathers is less than the average height of their fathers and the average height of sons of short fathers is more than the average height of their fathers. In other words the height of sons of tall fathers or short fathers will regress or go back to general average height. This phenomenon was described by him as 'regression' (to go back).

Lines of Regression

If the correlation between two variables is of a high degree then the points lie in a narrow strip. Then we can draw a line such that almost all the points should lie close to it from both sides. Such a line is called the *line of best fit*. It is called the *line of regression*.

Line of Regression of y on x :

$$y = ax + b.$$

This line is used for estimating the value of y given a value of x.

Line of Regression of x on y :

$$x = ay + b.$$

This line is used for estimating the value of x given a value of y.

Coefficients of Regression :

(i) Coefficient of regression of y on x : $b_{yx} = \dfrac{r\sigma_y}{\sigma_x}$

$$b_{yx} = \frac{\left(\Sigma \frac{xy}{N}\right) - \left(\Sigma \frac{x}{N}\right)\left(\Sigma \frac{y}{N}\right)}{\Sigma\left(\frac{x^2}{N}\right)\left(\Sigma \frac{x}{N}\right)^2}$$

The equation of the line of regression of y on x is given by $(y - \bar{y}) = b_{xy}\ (x - \bar{x})$.

(ii) Coefficient of regression of x on y : $b_{xy} = \dfrac{r\sigma_x}{\sigma_y}$

$$b_{xy} = \frac{\left(\Sigma \frac{xy}{N}\right) - \left(\Sigma \frac{x}{N}\right)\left(\Sigma \frac{y}{N}\right)}{\Sigma\left(\frac{y^2}{N}\right)\left(\Sigma \frac{y}{N}\right)^2}$$

The equation of the line of regression of y on x is given by $(x - \bar{x}) = b_{xy}\ (y - \bar{y})$.

Relation between Coefficient of Regression and Coefficient of Correlation

$$r^2 = b_{xy}\ .\ b_{yx}$$

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

The sign of r is dependent on the sign of b_{xy} and b_{yx}. If both are positive then r is positive, if both are negative then r is also negative.

Example 3.40

Find the equations of the lines of regression and also the coefficient of correlation from the following data:

x	62	64	65	69	70	71	72	74
y	126	125	139	145	165	152	180	208

Solution

x	y	x^2	y^2	xy
62	126	3844	15876	7812
64	125	4096	15625	8000
65	139	4225	19321	9035
69	145	4761	21025	10005
70	165	4900	27225	11550
71	152	5041	23104	10792
72	180	5184	32400	12960
74	208	5476	43264	15392

$$\bar{x} = \frac{547}{8} = 68.375$$

$$\bar{y} = \frac{1240}{8} = 155$$

$$b_{yx} = \frac{\left(\Sigma \frac{xy}{N}\right) - \left(\Sigma \frac{x}{N}\right)\left(\Sigma \frac{y}{N}\right)}{\Sigma\left(\frac{x^2}{N}\right)\left(\Sigma \frac{x}{N}\right)^2}$$

$$b_{yx} = \frac{\left(\frac{85546}{8}\right) - \left(\frac{547}{8}\right)\left(\frac{1240}{8}\right)}{\frac{37527}{8} - \left(\frac{547}{8}\right)^2}$$

$$b_{yx} = \frac{95.125}{705} = 6.0457$$

$$b_{xy} = \frac{\left(\frac{85546}{8}\right) - \left(\frac{547}{8}\right)\left(\frac{1240}{8}\right)}{\frac{197840}{8} - \left(\frac{1240}{8}\right)^2}$$

$$b_{xy} = \frac{95.125}{705} = 0.1349$$

$$r^2 = b_{yx}\ b_{xy} = 0.1349 \times 6.0457$$

$$r = \sqrt{b_{yx}\ b_{xy}} = 0.9031$$

The equation of the line of regression of yon x

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

$$y - 155 = 6.0457\ (x - 68.375)$$

i.e. $y = -258.3747 + 6.0457$

The equation of the line of regression of x on y

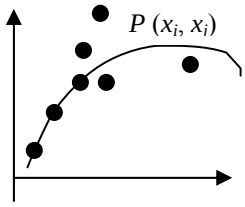
$$(x - \bar{x}) = b_{xy}\ (y - \bar{y})$$

$$x - 68.375 = 0.1349\ (y - 155)$$

i.e. $x = -47.4755 + 0.1349\,y$

Principle of Least Squares

Suppose we have a set of values (x_i, x_j). Suppose further that we want to fit a parabola $y = a + bx + cx^2$ to these values. The parabola must be close to the given points as much as possible. The principle of least squares states that the parabola should be such that the distances of the given points from the parabola measured along the y axis must be minimum.



Suppose $P(x_i, x_j)$ is a given point and a line parallel to the y axis intersects the curve $y = a + bx + cx^2$ in Q. Then Q is $(x_i, a + bx_i + cx_i^2)$.

We find a, b, c such that the distance PQ is minimum.

But distance PQ is minimum when the squares of the distance $(y_i - a - bx_i - cx_i^2)^2$ is minimum. This must be true for all points.

This means we should have $S = \Sigma\ (y_i - a - bx_i - cx_i^2)^2$ must be minimum. Since we find a, b, c such that the sum of the squares S is minimum, the method is known as the method of least squares.

Now for S to be minimum the conditions are

$$\frac{\partial s}{\partial a} = 0, \frac{\partial s}{\partial b} = 0, \frac{\partial s}{\partial c} = 0$$

$$\Sigma\ (y_i - a - bx_i - cx_i^2) = 0$$

$$\Sigma\ (y_i - a - bx_i - cx_i^2)\ x_i = 0$$

$$\Sigma\ (y_i - a - bx_i - cx_i^2)\ x_i^2 = 0$$

Thus we can write :

$$\Sigma y = Na + b\Sigma x + c\Sigma x^2$$

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

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

where N is the number of observations.

These equations are called normal equations.

If the curve to be fit is a straight line of the form $y = a + bx$ the normal equations are :

$$\Sigma y = Na + b\Sigma x$$

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

Fitting of the curve of the type $y = ab^x$ or $y = ax^b$

When it is required to fit a curve of the above type we take logarithms of both sides and then obtain normal equations as before.

Suppose $y = ab^x$

$$\log y = \log a + x \log b = A + xB$$

where $A = \log a$, $B = \log b$.

Proceeding as before we get the normal equations as :

$$\Sigma \log y = NA + B \Sigma x$$

$$\Sigma x \log y = A \Sigma x + B \Sigma x^2$$

Remark

If the data are equally spaced then we can change the variable x to u by the relation $u = x - A$ where A is the mid value. This makes $\Sigma u = 0$ and $\Sigma u^3 = 0$. This makes it easier to calculate the values of the constants.

Example 3.41

Fit a straight line curve to the following data taking x as the independent variable.

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.5

Solution

Let the straight line to be fit be $y = a + bx$. Then the normal equations are

$$\Sigma y = Na + b \Sigma x$$

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

x	y	xy	x^2
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.5	26	16
$\Sigma x = 10$	$\Sigma y = 17.1$	$\Sigma xy = 47.9$	$\Sigma x^2 = 30$

Hence the normal equations become :

$$= 5a + 10b \qquad \text{. . . (i)}$$

$$= 10a + 30b \qquad \text{. . . (ii)}$$

Solving the Eqs (i) and (ii)

From Eq. (i), we get,

$$= 10a + 20b$$

$$= 10a + 30b$$

$$= 10b \text{ thus } b = 1.37$$

$$5a = 17.1 - 10\,(1.37) = 3.4 \text{ thus } a = 0.68$$

The equation of the straight line is

$$y = 0.68 + 1.37x$$

Example 3.42

Fit a straight line to the following data

Year	1951	1961	1971	1981	1991
Production (in '000 tons)	10	12	8	10	13

Also estimate the production in 1987.

Solution

Since the values of x are odd and equispaced we change x to u by the relation $u = \dfrac{x - 1971}{10}$.

Let the straight line be $y = a + bu$. Then the normal equations are:

$$\Sigma y = Na + b \Sigma x$$

$$\Sigma uy = a \Sigma u + b \Sigma u^2$$

x	$u = \dfrac{x - 1971}{10}$	y	uy	u^2
1951	– 2	10	– 20	4
1961	– 1	12	– 12	1
1971	0	8	0	0
1981	1	10	10	1
1991	2	13	26	4
	$\Sigma u = 0$	$\Sigma y = 53$	$\Sigma uy = 4$	$\Sigma u^2 = 10$

Thus, the normal equations become :

$$53 = 5a \qquad a = 10.6$$

$$4 = 10b \qquad b = 0.4$$

Thus the equation of the straight line is $y = 10.6 + 0.4u$

But $\quad u = \dfrac{x - 1971}{10}$

$$y = 10.6 + (0.4)\left(u = \frac{x - 1971}{10}\right)$$

$$y = -68.24 + 0.04\,x$$

When $\quad x = 1987 \; y = -68.24 + 0.04\,(1987) = 11.24$

SAQ 11

(a) The following results were obtained from marks in Statistics and Mathematics in an examination :

	Marks in Statistics	**Marks in Mathematics**
Mean	47.5	39.5
Standard Deviation	16.8	10.8

$$r = 0.95$$

Find both the regression equations. Also estimate the value of y for $x = 30$.

(b) The equations of the two lines of regression are $6y = 5x + 90$ and $15x = 8y + 130$.

Find the means of x and y and the coefficient of correlation.

(c) Fit a second degree parabolic curve to the following data

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

(d) Find the equations of the lines of regression from the following data. Also find the coefficient of correlation r.

x	80	45	55	56	58	60	65	68	70	75	85
y	82	56	50	48	60	62	64	65	70	74	90

(e) The following data give the marks obtained by a class of students in two subjects. Find the lines of regression.

Paper I	56	55	58	58	57	56	60	54	59	57
Paper II	68	67	67	70	65	68	70	66	68	66

Calculate the coefficient of correlation, the two lines of regression and estimate the marks in Mathematics of a student who scores 62 marks in Statistics.

(f) Given the following values of x and y

x	3	5	6	8	9	11
y	2	3	4	6	5	8

Find the equation of regression of (i) y on x and (ii) x on y.

(g) The following table gives the age of cars of a certain make and annual maintenance costs. Obtain the regression equation for costs related to age.

Age of Cars (in years)	2	4	6	8
Maintenance Cost (in hundreds of Rs.)	10	20	25	30

3.16 SUMMARY

Let us summarise what we have learnt in this unit.

- The arithmetic of n individual observations $x_1, x_2, x_3, \ldots, x_n$ is given by (denoted by $\bar{x}$)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- The arithmetic mean $\bar{y}$ of a grouped data classified into k classes with class marks $y_1, y_2, \ldots, y_k$ and frequencies $f_1, f_2, \ldots, f_k$ is given by

$$\bar{y} = \frac{\sum_{i=1}^{k} f_i\, y_i}{\sum_{i=1}^{k} f_i} = a + \frac{\sum_{i=1}^{k} f_i\, d_i}{\sum_{i=1}^{k} f_i}$$

$$d_i = y_i - a$$

or $$\bar{y} = a + c\frac{\sum_{i=1}^{k} f_i\, u_i}{\sum_{i=1}^{k} f_i}$$

where $u_i = \frac{y_i - a}{c}$, c is the width of the class.

- The median M of n individual observations $x_1, x_2, \ldots, x_n$ is

$$M = \begin{cases} \left\{\frac{n+1}{2}^{\text{th}} \text{ observation}, \right. & \text{if } n \text{ is odd} \\ \dfrac{\frac{n}{2}\text{th observation} + \left(\frac{n}{2}+1\right)\text{th observation}}{2}, & \text{if } n \text{ is even} \end{cases}$$

- The median M of k observations $x_1, x_2, \ldots, x_k$ with frequencies $f_1, f_2, \ldots, f_k$ is given by

$$M = l + \frac{\frac{n}{2} - c_{f-1}}{f_m} \times c_m$$

where l = lower limit of the median class,

f_m = frequency of the median class,

c_{f-1} = commulative frequency of the class preceding the median class,

n = sum of all the frequencies, i.e. all the observations, and

c_m = width of the median class.

The median class being the class which contains the $\frac{n}{2}$th bservations.

- The mean deviation denoted by MD about median is

$$MD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{1}{n} \sum_{i=1}^{k} |d_i|,$$

n being the total number of individual observations, $x_1, x_2, \ldots, x_n$ and $\bar{x}$ being the median.

- In case of grouped data

$$\text{MD} = \frac{\sum_{i=1}^{k} f_i |(y_i - \bar{y})|}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i |d_i|}{\sum_{i=1}^{n} f_i}$$

where $d_i = y_i - \bar{y}$.

- The standard deviation denoted by SD or σ is given by

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

In case of grouped data

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i\ (y_i - \bar{y})^2}{\sum_{i=1}^{k} f_i}$$

- If $d_i = x_i - a$, where a is an assumed mean then

$$\sigma^2 = \frac{\sum_{i=1}^{n} d_i^2}{n} - \left(\frac{\sum_{i=1}^{n} d_i}{n}\right)^2$$

and in case of grouped data

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i\ d_i^2}{\sum_{i=1}^{k} f_i} - \left(\frac{\sum_{i=1}^{k} f_i\ d_i}{\sum_{i=1}^{k} f_i}\right)^2$$

If we define $u_i = \dfrac{y_i - a}{c}$, c being the class size, then

$$\sigma^2 = c^2 \left[\frac{\sum_{i=1}^{k} f_i\ u_i^2}{\sum_{i=1}^{k} f_i} - \left(\frac{\sum_{i=1}^{k} f_i\ u_i}{\sum_{i=1}^{k} f_i}\right)^2\right].$$

If we let A and B represent events and $P(A)$ and $P(B)$ the probabilities of their occurrence, then

(i) For events that are not mutually exclusive, the general additive rule is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

(ii) For mutually exclusive events, the additive rule is

$$P(A \text{ or } B) = P(A) + P(B)$$

(iii) If A and B are mutually exclusive, the probability of their joint occurrence P (A and B) is zero.

(iv) If A and B are exhaustive $P(A) + P(B) = 1.00$.

(v) If A and $\overline{A}$ are complementary, they are mutually exclusive [P (A and $\overline{A}$) = 0] and exhaustive [$P(A) + P(\overline{A}) = 1.00$].

(vi) If A and B are independent $P(A/B) = P(A)$.

(vii) If A and B are related or dependent $P(A/B) \neq P(A)$.

(viii) The multiplication rule for dependent event is $P(A \text{ and } B) = P(B)\, P(A/B)$.

(ix) The multiplication rule for independent event is P (A and B) = P (A) P (B).

(x) Boyes's theorem provides a basis for revising prior probabilities in the light of additional information or evidence by using the basic formula for conditional probability

$$P(A/B) = \frac{P(A \text{ and } B)}{P(B)}.$$

(xi) The Binomial Distribution $P(r) = {}^nC_r\, p^r\, q^{n-r}$.

(xii) The Poisson Distribution $P(r) = \dfrac{e^{-m}\, m^r}{r!}$.

(xiii) The Normal Distribution $f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

where, mean = μ, and variance = σ^2.

3.17 KEY WORDS

Sample Space : The set of all possible outcomes of an experiment is called sample space, denoted by S.

Event : An event is a subset of a sample space.

Exhaustive Events : The event $E_1, E_2, \ldots, E_n$ are called exhaustive if $E_1 \cup E_2 \cup \ldots \cup E_n = S$.

Mutually Exclusive Events : If two events cannot occur simultaneously, then they are called mutually exclusive events.

Union : The union of sets A and B is the set of all elements which belong either to A or to B or to both (i.e. the set of elements which belong to at least one of the sets A and B). This is written as $A \cup B$.

Intersection : The intersection of sets A and B is the set of all elements which belong to both A and B (i.e. the set of elements which are common to the sets A and B). This is written as $A \cap B$.

Complement : The complement of a set A is the set of all elements of the universal set S, which do not belong to A. This is denoted by A^c.

Disjoint Sets : Two sets A and B are said to be disjoint, if they have no common elements (i.e. not a single element of B is contained in A, and none of A is contained in B). If A and B are disjoint sets then
$A \cap B = \phi$.

3.18 ANSWERS TO SAQs

SAQ 1

(a) 38.9 paise.

(b) 8.8 years, 8.2 years.

(c) 40.617 mm.

(d) Median = 23.5, 50% of the student obtain less than 23.5 marks out of 50.

SAQ 2

(a) Median = 24.5.

(b) 17.2 units, 17.5 units.

(c) 3.75.

(d) 636.1.

SAQ 3

(a) SD = 7.26 nearly.

(b) 4.5, 31.85.

(c) $\frac{54}{11}$

SAQ 4

(a) (i) The probability that the student will pass at least one course

$$= 0.6 \times 0.3 + 0.7 \times 0.4 + 0.6 \times 0.7$$
$$= 0.88$$

(ii) The probability that the student will fail both courses

$$= 0.4 \times 0.3$$
$$= 0.12$$

(b) Reliability of the system

$$= 1 - 0.2 \times 0.2$$
$$= 0.96 \quad \text{or} \quad 96\%$$

(c)

Boys	Girls	Poor	Rich	Fair	Not Fair
55	25	70	10	20	60

Required probability = $\dfrac{20 \times 10 \times 25}{80 \times 80 \times 80} + \dfrac{70 \times 55 \times 60}{80 \times 80 \times 80} = \dfrac{59}{128}$

(d) They will say the same when they both say true or both say false. So, required probability

$= 0.7 \times 0.8 + 0.3 \times 0.2 = 0.62$

(e) The required probability = $\dfrac{0.4 \times 0.6}{0.4 \times 0.6 + 0.7 \times 0.4} = \dfrac{6}{13}$.

(f) The required probability

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

$$= 1 - 0.6 \times 0.7 \times 0.8 \times 0.9$$

$$= 1 - 0.3024$$

$$= 0.6976\,.$$

SAQ 5

(a) E_1 be the event when the lot contains exactly 2 defective articles and E_2 be the event when it contains exactly 3 defective articles

$$P\,(E_1) + P\,(E_2) = 0.4 + 0.6 = 1.0$$

E be the event that represents the 12th draw.

According to the theorem of total probability for compound events,

We have $\quad P(E) = P(E_1)\, P\left(\dfrac{E}{E_1}\right) + P\,(E_2)\, P\left(\dfrac{E}{E_2}\right)$

To end the testing at the 12th draw from a sample of 20 articles containing
2 defective articles, we must have for $\boldsymbol{P\,(E\,/\,E_1)}$: The first 11 draws must contain 1 defective and 10 non-defective articles and the 12th draw must give a defective article.

The probability of these events are :

$$\frac{{}^{18}C_{10} \times {}^{2}C_1}{{}^{20}C_{11}} \text{ and } \frac{1}{9} \text{ respectively .}$$

Hence $\quad P\left(\dfrac{E}{E_1}\right) = \dfrac{{}^{18}C_{10} \times {}^{2}C_1}{{}^{20}C_{11}} \times \dfrac{1}{9} = \dfrac{11}{190}$

For $\boldsymbol{P\,(E\,/\,E_2)}$: The first 11 draws must contain 2 defective and 9 non-defective articles and the 12th draw must give a defective article. The probability of these events are :

$$\frac{{}^{17}C_9 \times {}^{3}C_2}{{}^{20}C_{11}} \text{ and } \frac{1}{9} \text{ respectively}$$

$$P\left(\frac{E}{E_2}\right) = \frac{{}^{17}C_9 \times {}^{3}C_2}{{}^{20}C_{11}} \times \frac{1}{9} = \frac{11}{228}$$

Now, we get, after substituting;

$$P\,(E) = 0.4 \times \frac{11}{190} + 0.6 \times \frac{11}{228} = \frac{99}{1900}$$

(b) The sequence of balls drawn is 2 black, 4 white, and 3 red.

Lets $P_1, P_2, \ldots, P_9$ denote the probability of drawing black, black, white, white, white, white, red, red, and red balls respectively,

The required probability, $P = P_1 \times P_2 \times P_3 \times \ldots \times P_8 \times P_9$.

So, the required probability,

$$P = \frac{2}{9} \times \frac{1}{8} \times \frac{4}{7} \times \frac{3}{6} \times \frac{2}{5} \times \frac{1}{4} \times 1 \times 1 \times 1 = \frac{1}{1260}$$

(c) Probability of 2 white balls are withdrawn from 2 white balls and 3 non-white balls = $\frac{{}^2C_2}{{}^5C_2} = \frac{1}{10}$.

Similarly, $\frac{{}^3C_2}{{}^5C_2} = \frac{3}{10}$; $\frac{{}^4C_2}{{}^5C_2} = \frac{3}{5}$; and $\frac{{}^5C_2}{{}^5C_2} = 1$

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

(d) (i) Required probability with replacement $= \frac{10}{25} \times \frac{15}{25} = 0.24$

(ii) Required probability without replacement $= \frac{10}{25} \times \frac{15}{24} = 0.25$.

SAQ 6

(a) The probability that the defective bolt comes from

$$\text{Machine A} = \frac{0.05 \times 0.25}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.40} = 0.362$$

$$\text{From machine B} = \frac{0.04 \times 0.35}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.40} = 0.406$$

Similarly, from machine $C = 0.232$

(b) The probability that the defective unit comes from Machine D

$$= \frac{0.03 \times 0.35}{0.01 \times 0.15 + 0.02 \times 0.20 + 0.02 \times 0.30 + 0.03 \times 0.35} = 0.4772$$

SAQ 7

(a) Here $p = 0.2$, $q = 1 - p = 0.8$

$n = 4$

(i) We know for binomial distribution

$$P(r) = {}^nC_r \; p^r \; q^{n-r}$$

Here $r = 1$

$$P(1) = {}^4C_1 \; 0.2^1 \; 0.8^3$$

$$= 0.4096$$

(ii) $P(0) = {}^4C_0\ 0.2^0\ 0.8^4$

$= 0.4096$

(iii) Atmost 2

The required probability

$= P(0) + P(1) + P(2)$

$= {}^4C_0\ 0.2^0\ 0.8^4 + {}^4C_1\ 0.2^1\ 0.8^3 + {}^4C_2\ 0.2^2\ 0.8^2$

$= 0.9728$

(b) We know; $\sum_{i=0}^{6} p(x_i) = 1$

So, $k + 3k + 5k + 7k + 9k + 11k + 13k = 1$

$\Rightarrow \quad k = \frac{1}{49}$

(i) Now, $P(X < 4) = 16k = \frac{16}{49}$

$P(X \geq 5) = 24k = \frac{24}{49}$

Now, $P(3 < X \leq 6) = 33k = \frac{33}{49}$

(ii) $P(X \leq 2) = 9k > 3$

$\therefore \quad k > \frac{1}{3}$

Thus, the minimum value of $k = \frac{1}{3}$.

(c) We know that, $\sum p(x_i) = 1$

$\Rightarrow \quad 0.1 + k + 0.2 + 2k + 0.3 + k = 1$

$\Rightarrow \quad k = 0.1$

Mean, $E(X) = \sum x_i\ p(x_i) = 0.8$

Variance = σ^2

$= \sum (x_i - \mu)^2\ p(x_i) = 2.16$

(d) (i) Since the drug is effective 40% of the time

So, $p = 0.4$

$q = 1 - p = 0.6$

$P(r) = {}^nC_r\ p^r\ q^{n-r}$

Now, $P(2) = {}^4C_2\ 0.4^2 \times 0.6^2 = 0.3456$

(ii) We know, mean $m = np$ and standard deviation $\sigma = \sqrt{npq}$.

Here, $p = 0.05$, $q = 0.95$, and $n = 60$

Mean, $m = 60 \times 0.5 = 3$

and S. D., $\sigma = \sqrt{60 \times 0.05 \times 0.95} = 1.688$

(e) The problem is a binomial one where probability of r number of successes out of n number of trials,

$$P(r) = {}^nC_r \; p^r \; q^{n-r}$$

Here, $n = 6$, $p = \frac{1}{3}$, and $q = \frac{2}{3}$

(i) $P(4) = {}^6C_4 \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^2 = 0.0825$

(ii) No correct answer, so that $r = 0$

$$P(0) = {}^6C_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^6 = 0.088$$

(iii) Atmost 2 correct answer,

Hence

$$P(0, 1, 2) = {}^6C_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^6 + {}^6C_1 \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^5 + {}^6C_2 \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^4 = 0.68$$

SAQ 8

(a) Here $p = \frac{1}{2400}$, $n = 200$

$m = n\,p$

$$= \frac{1}{2400} \times 200 = \frac{1}{12}$$

We know $P(r) = \frac{e^{-m}\, m^r}{r!}$

The required probability $= 1 - \frac{e^{-m}\, m^0}{0!} = 0.08$

(b) Here $p = \frac{1}{500}$, $n = 10$

So, $m = n\,p$

$$= \frac{1}{50} = 0.02$$

We know $P(r) = \frac{e^{-m}\, m^r}{r!}$

For no defective probability, $P(0) = \frac{e^{-m}\, m^0}{0!} = 0.9802$

So, the no. of lots = 0.9802 × 10000 = 9802

For one defective;

$$P(1) = \frac{e^{-0.02}\ 0.02^1}{1!} = 0.0196$$

So, the no. of lots = 0.0196 × 10000 = 196

For two defective,

$$P(2) = \frac{e^{-0.02}\ 0.02^2}{2!}$$

$$= 1.96 \times 10^{-4}$$

So, no. of lots $= 1.96 \times 10^{-4} \times 10000$

$= 1.96$

≈ 2

(c) Poisson Distribution is found to be particularly useful in waiting line.

We know, $P(r) = \dfrac{e^{-m}\, m^r}{r!}$

Here $m = 2$,

(i) when $r = 0$, $P(0) = \dfrac{e^{-2}\ 2^0}{0!} = 0.1353$

(ii) when $r = 1$, $P(1) = \dfrac{e^{-2}\ 2^1}{1!} = 0.2707$

(iii) when $r = 2$, $P(2) = \dfrac{e^{-2}\ 2^2}{2!} = 0.2707$

Then, the probability of more than two arrivals in 10 minute :

$$P(x > 2) = 1 - [P(0) + P(1) + P(2)]$$

$$= -\ (0.1353 + 0.2707 + 0.2707)$$

$$= 0.3233$$

SAQ 9

(a) Now, $z = \dfrac{x - \mu}{\sigma} = 0$

Area lying to the right of $z = 0$ is 0.50. So, % of students getting more than 60 marks is 50%.

(b) The normal curve with necessary information is given in Figure 11.3.

Since 11.51% of the soldiers were taller than X_2 = 70.4 inch, then 38.49% of the soldiers must have height between the value of the mean (μ) and X_2, because the total area must add up to 50%. Similarly, the area between the mean and X_1 = 65.4 inch is 40.32%. Given these areas, we can find the value of z with respect to area 40.32% and the value of z against the area of 38.49%.

From z table, $z_1 = 1.3$, and $z_2 = 1.2$.

Figure 3.6 : The Normal Curve with Necessary Information

Knowing that $X_2 = \mu + z_2\,\sigma$, and $X_1 = \mu - z_1\,\sigma$ considering the following equation

$$70.4 = \mu + 1.2\,\sigma$$

and $\quad 65.4 = \mu - 1.3\,\sigma$

After solving, we get

$$\sigma = 2; \quad \mu = 68 \text{ inches.}$$

SAQ 10

(b)

x	x^2	y	y^2	xy
2.6	6.76	5.9	34.81	15.34
2.8	7.84	6.0	36	16.80
2.9	8.41	6.2	38.44	17.98
3.1	9.61	6.2	38.44	19.22
3.2	10.24	7.6	57.76	24.32
2.3	5.29	7.0	49	16.10
2.5	6.25	7.5	56.25	18.75
1.8	3.24	5.5	30.25	9.90
Σ 21.2	Σ 57.64	Σ 51.9	Σ 340.95	Σ 138.41

$$r = \frac{\left(\Sigma \frac{xy}{n}\right) - \left(\Sigma \frac{x}{n}\right)\left(\Sigma \frac{y}{n}\right)}{\sqrt{\Sigma\left(\frac{x^2}{n}\right) - \Sigma\left(\frac{x}{n}\right)^2 - \Sigma\left(\frac{y^2}{n}\right) - \Sigma\left(\frac{y}{n}\right)^2}}$$

$$= \frac{0.1094}{0.3113} = 0.3514.$$

SAQ 11

(a) The equation of the line of regression of y on x is $(y - \bar{y}) = b_{yx}\ (x - \bar{x})$

$$b_{yx} = \frac{r\sigma_x}{\sigma_y}$$

$$y - 39.5 = 0.95\left(\frac{10.8}{16.8}\right)(x - 47.5)$$

thus $y = 10.49 + 0.6107\ x$

when $x = 30,\ y = 28.8$

The equation of the line of regression of x on y $(x - \bar{x}) = b_{xy}\ (y - \bar{y})$

$$b_{xy} = \frac{r\sigma_x}{\sigma_y}$$

$$x - 47.5 = 0.95\left(\frac{16.8}{10.8}\right)(y - 39.5)$$

thus $x = -10.87\ + 1.4778\ y$

(b) We solve both the equations simultaneously

$$6y = 5x + 90$$

$$15x = 8y + 130$$

$$10y = 400 \text{ thus } y = 40$$

Substituting this value of y in the first equation gives us

$$240 = 5x + 90 \text{ thus } x = 30$$

The two lines of regression intersect at $\bar{x}$ and $\bar{y}$, thus $\bar{x} = 30,\ \bar{y} = 40$

To find r we rewrite the two lines of regression as :

$$y = \frac{5}{6}x + 15$$

thus $b_x = \frac{5}{6}$

$$x = \frac{8}{15} + \frac{130}{15}$$

thus $b_y = \frac{8}{15}$

$$r = \sqrt{b_{yx}\ b_{xy}} = \sqrt{\frac{5}{6}\cdot\frac{8}{15}} = \frac{2}{3}$$

(c) Since the values of x are odd and are equally spaced we change x to u by the relation $u = x - 5$.

Let the equation of the parabola be

$$y = a + bu + cu^2$$

The normal equations are :

$$\Sigma y = Na + b \Sigma u + c \Sigma u^2$$

$$\Sigma uy = a \Sigma u + b \Sigma u^2 + c \Sigma u^3$$

$$\Sigma u^2 y = a \Sigma u^2 + b \Sigma u^3 + c \Sigma u^4$$

x	u	y	u^2	u^3	u^4	uy	u^2y
1	– 4	2	16	– 64	256	– 8	32
2	– 3	6	9	– 27	81	– 18	54
3	– 2	7	4	– 8	16	– 14	28
4	– 1	8	1	– 1	1	– 8	8
5	0	10	0	0	0	0	0
6	1	11	1	1	1	11	11
7	2	11	4	8	16	22	44
8	3	10	9	27	81	30	90
9	4	9	16	64	256	36	144
	$\Sigma u = 0$	**$\Sigma u = 74$**	**$\Sigma u^2 = 6\ 0$**	**$\Sigma u^3 = 0$**	**$\Sigma u^4 = 708$**	**$\Sigma uy = 51$**	**$\Sigma u^4 y = 411$**

Now the normal equations become :

$$74 = 9\,a + 60\,c \qquad \ldots (1)$$

$$51 = 60\,b \quad b = 0.85 \qquad \ldots (2)$$

$$411 = 60\,a + 708\,c \qquad \ldots (3)$$

To find a and c multiply (1) by 60 and (3) by 9

$$4440 = 540\,a + 3600\,c$$

$$3699 = 540\,a + 6372\,c$$

$$741 = -2772\,c \quad \text{thus} \quad c = -0.2673$$

$$9\,a = 74 - 60\,c = 74 + 16.038 = 90.038$$

$$a = 10.0042$$

Thus the equation of the parabola is

$$y = 10 + 0.85\,u - 0.27\,u^2 = 10 + 0.85\,(x - 5) - 0.37\,(x - 5)^2$$

$$y = -1 + 3.55\,x - 0.27\,x^2$$

(d) ($y = 0.9044 + 0.9917\,x$, $x = 9.3810 + 0.8513\,y$, $r = 0.9188$).

(e) ($y = 39 + 0.5\,x$, $x = 15.67 + 0.6122\,y$, $r = 0.5533$, $y = 70$).

(f) ($y = 0.7143\,x - 0.3334$; $x = 1.2857\,y + 1.0001$).

(g) (x : age; y : cost; $y = 3.25\,x + 5$).

FURTHER READINGS

Shanti Narayan, *Integral Calculus*, by S. Chand & Co.

Georgy Thomas and Ross L. Finney, *Calculus and Analytic Geometry*, Narosa Publishing House.

M. K. Singal and Asha Rani Singal, *Complete Course in I. S. C. Mathematics Part-I and Part-II*, Pitambar Publishing House.

APPLIED MATHEMATICS

This course on Applied Mathematics (BNA-011) is in your hand. The whole course is comprises three blocks : Algebra and Statistics, Solid Geometry and Calculus.

In first block, you will be introduced to basic concepts of algebra and statistics. The concept of complex numbers and its properties has been explained. The important De-Moiver's theorem has also been discussed. The usefulness of vectors in engineering and mathematics results from the fact that many physical quantities – for example, forces and velocities – may be represented by vectors, and in several respects the rules of vector calculations are as simple as the rules governing the system of real numbers. However, vector analysis is a shorthand which simplified may calculations considerably them. For all these reasons, extensive use is made of vector notation in modern engineering literature. The basic concepts of vectors, the different operations on vectors and the vector products of two vectors have also been discussed. Probability and statistics have relevance in our lives, which have been introduced in this block.

Block 2 covers co-ordinate Geometry, 3-Dimensional Geometry and Spherical Trigonometry. You will be introduced to standard and general equation of circles, parabola, ellipse and hyperbola. In 3-dimensional geometry, the sphere cones and cylinders will be discussed. You will also be introduced to spherical trigonometry.

In Block 3, you will be introduced to basic concepts of differential and integral calculus. This includes limit, continuity, mean value theorem and application of calculus in engineering field. The matter of different and integral calculus is presented in single block because one has to continuously refer to the results of differential calculus while studying integral calculus.

We believe that the course on applied mathematics will provide insights, which help you in understanding other subjects in Diploma in Nautical Science.

ALGEBRA AND STATISTICS

The knowledge and concepts of mathematics help in handling any engineering problem. In first block of Applied Mathematics, which is in your hands, you will be studying some concepts of mathematics, which are a must for you. This block will be handling with the concepts of algebra, vectors and probability and statistics.

Unit 1 deals with complex numbers; Cartesian, polar and exponential forms. De-Moivre's theorem. Determinants and its properties has also been explained in this unit. You will also be introduced to infinite series; convergence of infinite series, uniform convergence and its properties.

Vectors are used extensively in almost all branches of physics, mathematics and engineering. In Unit 2, we have presented the basic concepts of vectors. The different operations on vectors and the vector products of two vectors have been discussed.

Unit 3, gives an introduction to probability. Probability is one of our most important conceptual tools, because we use it to assess degrees of uncertainty and thereby to reduce risk. Whether or not you have had formal instruction in this topic, you are already familiar with the concept of probability since it pervades almost all aspects of our lives without consciously realizing it, many of your decisions are based on probability. For example, when you study for an examination, you concentrate more on areas that you feel are likely to be covered in the test. You may cancel or postpone an outdoor activity. If you believe that the likelihood of rain is high.

For want of clarity in concepts, number of solved examples have been introduced in each unit. To help you check your understanding and to assess yourself, each unit contains SAQs. The answers to these SAQs are given at them only after attempting the exercises.

At the end, we wish you all the best for your all educational endeavors.