BACHELOR OF COMMERCE

BC 402

BUSINESS STATISTCS-II

Directorate of Distance Education
Guru Jambheshwar University of Science & Technology, Hisar – 125001

CONTENTS

Lesson No: 1	**Author:** Dr. B. S. Bodla
Updated By: Ms. Chand Kiran	**Vetter:** Dr. Ajay Suneja

LESSON 1

CORRELATION

STRUCTURE

1.0 LEARNING OBJECTIVES

After reading this lesson, you should be able to-

- Explain meaning and types of correlation;
- Understand and determine correlation with the help of various formulas; and
- Explain applications of Karl Pearson's Coefficient of Correlation.

1.1 INTRODUCTION

In data analysis, the concept and technique of correlation is of great importance. The fundamentals of this concept were first, propounded by the French astronomer, Bravais, but

correlation technique as such was first investigated graphically by Sir Francis Galton. In 1896, Karl Pearson introduced a method of assessing correlation by means of coefficient of correlation. These two statisticians (Galton and Pearson) studied many problems of Biology and Genetics with the help of this technique. In economics, this technique has special uses. Regarding utility of correlation in Economics, W.A. Neiswanger writers, "Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective". The concepts of regression and ratio of variation are based upon the measure of correlation. Measure of correlation ensures that prediction of variables will be reliable. In the words of Tippett, "the effect of correlation is to reduce the range of uncertainty of our prediction". The prediction based on correlation analysis will be more reliable and near to the reality.

1.2 MEANING OF CORRELATION

Correlation refers to the relationship of variables. Some relationship is found in certain type of variables, for example there exists a relationship between price and demand, production and employment, wages and price index, etc. Even everyday experience demonstrates, how far different phenomena are related to each other in some way, for example, it is found that there is close relationship between ages of husbands and wives, stature of fathers and sons, heights and weights of young men, capital invested and profits earned and so on. Correlation is a statistical technique which measures and analyses the degree or extent to which two variables or phenomena fluctuate with reference to each other. Croxton and Cowden state that, "when the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation".

Connor states that, "If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other then they are said to be correlated". According to Prof. Boddington, "whenever some definite connection exists between the two or more groups, classes or series of data, they are said to be correlation". According to E. Davenport, "the whole subject of correlation refers to that inter-relation between separate character by which they tend,

in some degree atleast, to move together". Thus, correlation studies the inter-dependence or relationship between variables. It is a statistical technique that measures the nature and extent of relationship between two or more variables.

1.2.1 CORRELATION AND CAUSATION

The existence of correlation between two (or more) variables only implies that these variables (i) either tend to increase or decrease together or (ii) an increase (or decrease) in one is accompanied by the corresponding decrease (or increase) in the other. The questions of the type, whether changes in a variable are due to changes in the other, are not answered by the study of correlation analysis. If there is a correlation between two variables, it may be due to any of the following situations:

(I) ONE OF VARIABLE MAY BE AFFECTING THE OTHER

A correlation coefficient calculated from the data on quantity demanded and corresponding price of tea would only reveal that the degree of association between them is very high. It will not give us an idea about whether price is affecting demand of tea or vice-versa. In order to know this, we need to have some additional information apart from the study of correlation. For example if, on the basis of some additional information, we say that the price of tea affects its demand, the price will be the cause and quantity will be the effect. The casual variable is also termed as independent variable while the other variable is termed as dependent variable.

(II) THE TWO VARIABLES MAY ACT UPON EACH OTHER

Cause and effect relation exits in this case also but it may be very difficult to find out which of the two variables is independent. For example, if we have data on price of wheat and its cost of production, the correlation between them may be very high because higher price of wheat may attract farmers to produce more wheat and more production of wheat may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of wheat. For the purpose of determining a relationship between the two variables in such situations, we can take any one of them as independent variable.

(III) THE TWO VARIABLES MAY BE DEPENDENT UPON BY OUTSIDE INFLUENCES

In this case we might get a high value of correlation between the two variables, however, apparently no cause and effect type relation seems to exist between them. For example, the demands of the two commodities, say X and Y, may be positively correlated because the incomes of the consumers are rising. Coefficient of correlation obtained in such a situation is called a spurious or nonsense correlation.

(IV) A HIGH VALUE OF THE CORRELATION COEFFICIENT MAY BE OBTAINED DUE TO SHEER COINCIDENCE (OR PURE CHANCE)

Given the data on any two variables, one may obtain a high value of correlation coefficient when in fact they do not have any relationship. For example, a high value of correlation coefficient may be obtained between the size of shoe and the income of persons of a locality. This is another situation of spurious correlation.

1.3 TYPES OF CORRELATION

The correlation may be of following types:

(1) Positive or Negative

(2) Simple, Partial or Multiple

(3) Linear or Nonlinear

1. Positive or Negative Correlation: Positive or direct correlation refers to the movement of the variables in the same direction. As one variable increases the other also increases or as one decreases, the other also decreases. This sort of correlation exists between supply and price of a commodity. For example, when price of a commodity increases, the supply of that commodity also increases or vice versa. 'Negative or 'inverse' correlation refers to, when one variable increases or decreases, the other moves in the reverse direction. Such a correlation is found between price and demand.

The following are the examples of Positive and negative correlation:

Positive Correlation		Negative Correlation	
(a) Both variables	(b) Both variables	(a) One variable	(b) One variable

X	Y	X	Y	X	Y	X	Y
70	35	27	44	60	42	27	25
75	38	23	42	65	38	23	32
80	43	290	39	72	33	20	36
90	50	19	36	77	29	19	39
110	56	17	32	82	25	17	42
120	62	13	25	86	18	13	44

2. Simple, Partial or Multiple Correlation. When only two variables are involved, the analysis of relationship between them is described as simple correlation. When more than two variables are involved, for example when there are three or four variables, and they are to be studied in relation to their relationship with one another, it is called multiple correlation. For example, when we study relationship between, say, agricultural production, rainfall and the amount of fertilizes used, it will be case of multiple correlation. In partial correlation, the relationship of two variables is studied by eliminating the effect of other variables from both.

3. Linear or Nonlinear Correlation : If the ratio of change between two variables is uniform then there will be linear correlation between them. Their relationship is best described by a straight line. If the variables under study are graphed, the plotted points will form a straight line. In nonlinear relationship, which can also be said curvilinear, the amount of change in one variable does not bear a constant ratio to the amount of change in the other variables. The following data show this phenomena :

Linear Correlation Curvilinear Correlation

X	Y	X	Y
20	50	50	10
40	100	55	12
60	150	60	22
80	200	90	34
100	250	98	45
120	300	120	56

It may be mentioned here that in practice, we find curvilinear relationship in most of the phenomena. However, since the technique of nonlinear correlation analysis being very complicated one, we generally assume that the relationship between the variables under study is linear.

1.4 METHODS OF DETERMINING CORRELATION

The methods of finding out correlation are:

(i) Scatter Diagram

(ii) Karl Pearson's Coefficient of Correlation,

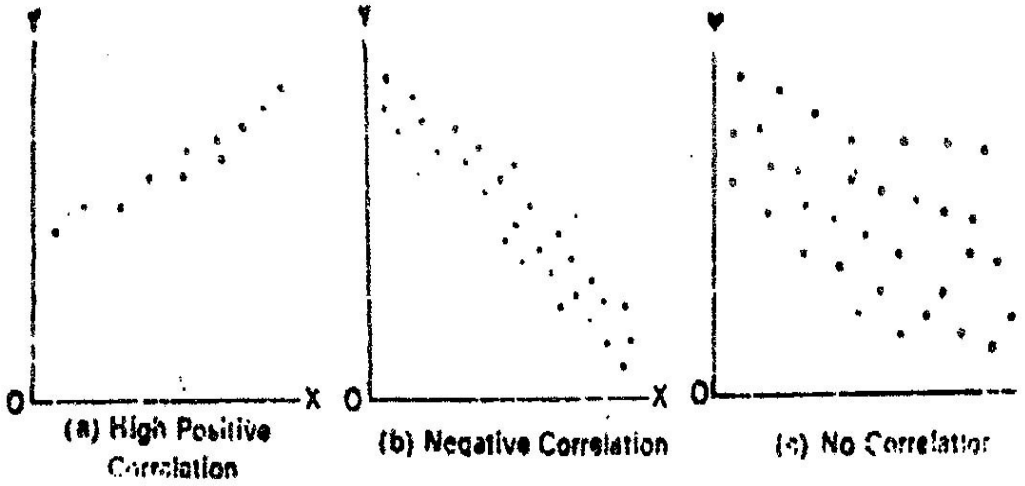
(iii) Spearman's Rank Coefficient of Correlation,

(iv) Coefficient of Concurrent Deviations.

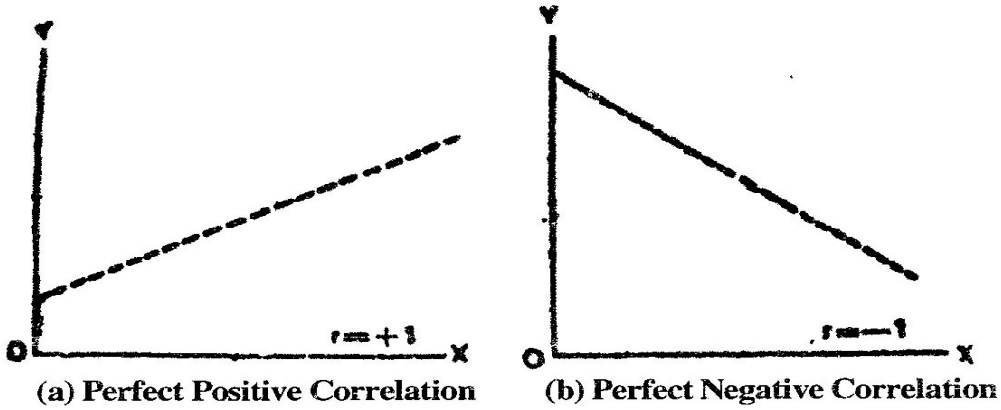
(v) Least Square method

(I) SCATTER DIAGRAM

One can get some idea whether there is any relationship present in two variables by plotting the values on a chart known as 'scatter diagram'. The chart is prepared by measuring X-variable on the horizontal axis and the Y-variable on the vertical axis and plot a point for each pair of observation of X and Y values. In this way the whole data are plotted in the shape of points. If these points show trend either upward or downward, the two variables are correlated. If the plotted points do not show any trend the two variables have no correlation. If the trend of the points is upward rising from left bottom and going up towards the right top, correlation is positive. On the other hand, if the tendency is reverse so that the points show a downward trend from the left top to the right bottom correlation is negative. The scatter diagram will take the following shapes:



If the points take the shape of a line, and it rises from left bottom going up toward the right top, then there is perfect positive correlation between the two variables. If the line moves in the reverse way, then there is perfect negative correlation between the two variables. This is shown below:



Scatter diagram is a simple and attractive method of ascertaining the nature of correlation between two variables. At a glance one can know whether variables are correlated or not, and if they are correlated whether correlation is positive or negative.

One of the serious limitations of scatter diagram is that the degree of correlation can't be known by this method. It gives only a rough idea of how the two variables are related, but definite conclusions cannot be drawn by merely examining the diagram.

(II) KARL PEARSON'S COEFFICIENT OF CORRELATION

A widely used measure of the degree of relationship between two variables is the coefficient of correlation, which is represented by the symbol 'r'. The noted statistician, Karl Pearson, who developed much of the theory of correlation analysis in the latter part of the last century, propounded this formula for calculating coefficient of correlation. Therefore, it is named after him, and is called 'Pearsonian coefficient of correlation' (also referred to as product-moment coefficient). Karl Pearson's formula is based on the following assumptions:

(1) The two series sought to be correlated are affected by a large number of independent causes which bring about a normal distribution in the series. (2) The forces affecting the distribution of items in the two series are related to each other in a relationship of cause and effect. (3) There is linear relationship between both the series. This means that if points are plotted on a scatter diagram the plotted points will fall all along a line.

Main features of Karl Pearson's coefficient of correlation are:

1. **Indication of the direction:** The algebraic sign of Pearsonian coefficient of correlation indicates the direction of relationship between two variables. If coefficient is in plus (+), it will be positive correlation, in case of minus (–) the correlation is negative.

2. **Indication of the Degree:** Karl Pearson's coefficient of correlation is a quantitative measure of relationship between two variables, and it lies between $\pm$ 1($-1 \leq r \leq +1$). Plus one (+1) represents perfect positive correlation and minus one (-1) perfect negative correlation. Zero represents absence of correlation.

The results between $\pm$1 are interpreted as having correlation of different degrees, based on how far the coefficient is way from zero and near one.

3. **Covariance:** Karl Pearon's coefficient of correlation is based on covarianace. Covariance is the arithmetic mean of cross-products, i.e., $(X - \overline{X})$, $(Y - \overline{Y})$ or $d_x d_y$. Thus, the formula for finding covariance is :

Covariance = $\frac{\sum(X-\overline{X})(Y-\overline{Y})}{N}$ Or $\frac{\sum d_x d_y}{N}$

The coefficient of correlation is independent of the change of origin and scale of measurements. In order to prove this property, we change origin and scale of both the variables X and Y:

Let $u_1 = \frac{X_i - A}{h}$

and $v_1 = \frac{Y_i - B}{k}$ where the constants A and B

refer to change of origin and the constants h and k refer to change of scale. We can write

$$X_i = A + hu_1 \quad \therefore \quad X = A + hu$$

Thus, we have $X_i - \bar{X} = A + hu_1 - A - hu = h(u_1 - u)$

Similarly, $Y_i = B + kv_i, \quad \therefore \quad Y = B + kv$

Thus, $Y_i - Y + kv_1 - B - kv = k(v_1 - v)$

The formula for the coefficient of correlation beween X and Y is

$$r_{XY} \frac{\Sigma(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2}\sqrt{\Sigma(Y_i - \bar{Y})^2}}$$

Substituting the values of $(X_i - \bar{X})$ **and** $(Y_i - \bar{Y})$ **we get**

$$r_{XY} = \frac{\Sigma h(u_t - u)\, k(v - v)}{\sqrt{\Sigma h^2(u_i - u)^2}\sqrt{\Sigma k^2(u_i - u)^2}} = \frac{\Sigma h(u_i - u)(v_i - v)}{\sqrt{\Sigma(u_i - \bar{u})^2}\sqrt{\Sigma(u_i - \bar{u})^2}}$$

$$\therefore \qquad r_{XY} = r_{uv}$$

This shows that correlation between X and Y is equal to correlation between u and v, where u and v are the variables obtained by change of origin and scale of the variables X and Y respectively.

DIRECT METHOD:

To calculate Karl Pearson's coefficient of correlation of individual series the following process is involved:

(i) Arithmetic means of both the series (X) and (Y) are calculated.

(ii) Deviation of the X-series and Y-series are computed from their respective artithmetic

means, for all individual values. The deviations are denoted by symbols d_X and d_Y :

$$d_x = (\bar{X}-X),\ d_v = (\bar{Y}-Y)$$

(iii) The deviations are squared and added to find out Σd^2_x and Σd^2_y

(iv) The individual d_x and d_y values are multiplied ($d_x d_y$) and their total ($\Sigma d_x d_y$) is obtained.

(v) Standard deviation of both the series are computed from the following formula :

$$\sigma_x = \sqrt{\frac{\Sigma D^2_x}{N}} \qquad \sigma_v = \sqrt{\frac{\Sigma D^2_v}{N}}$$

(vi) The coefficient of correlation is obtained from the following formula :

$$r = \frac{\Sigma d_x d_y}{N\sigma_x \sigma_v}$$

or

$$\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N\sigma_x \sigma_v}$$

or

$$\frac{1}{N}\ \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sigma_x \sigma_v}$$

Where,
r = Pearsonian coefficient of correlation
$\Sigma d_x d_y$ = Total of the products of the deviation of values from the respective arithmetic means in both the series
N = Numbr of items
σ_x = Standard Deviation of X-series
σ_y = Standard Deviation of Y-series

This direct method can further be simplified if instead of calculating standard deviations of the series, the various values are taken directly to the formula for coefficient of correlation. The formula for calculating 'r' can thus be written as:

$$r = \frac{\Sigma d_x d_y}{N\sqrt{\frac{\Sigma d^2_x}{N} \times \frac{\Sigma d_y{}^2}{N}}} \quad \text{or} \quad \frac{\Sigma d_x d_y}{\sqrt{\Sigma d^2_x \times \Sigma d^2_y}}$$

Production Moment Correlation: If there is a small number of items in the two series, their correlation can be found out by the product moment method. The formula for such method is:

$$r = \frac{A\sum XY - XY}{\sqrt{(A\sum X^2 - X^2) \times (A\sum Y^2 - Y^2)}}$$

Where:

$A\sum XY$ represents the arithmetic means of the summation of the products of X and Y.

X and Y represent arithemtic average of X and Y series

$A\sum X^2$ and $A\sum Y^2$ represent arithmetic means of summation of squares of the items of X and Y series respectively.

Illustration 1: Calculate the Karl Pearson's coefficient of correlation from the following pairs of values:

Values of X :	12	9	8	10	11	13	7
Values of Y :	14	8	6	9	11	12	3

SOLUTION:

The formula for Karl Pearson's coefficient of correlation is

$$r_{XY} = \frac{N\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

The values of different terms, given in the formula, are calculated from the following table:

X_i	Y_t	$X_t Y_t$	X^2_i	$X^2 f$
12	14	168	144	196
9	8	72	81	64
8	6	48	64	36
10	9	90	100	81

11	11	121	121	121
13	12	156	169	144
7	3	21	49	9
70	63	676	728	651

Here n=7 (no. of pairs of observations)

$$r_{XY} = \frac{7\times676-70\times63}{\sqrt{7\times728 - (70)^2}\sqrt{7\times651 - (63)^2}}$$

$$= \frac{322}{14\times24.24} = 0.94 \quad -$$

Illustration 2 : Calculate the Karl Pearson's coefficient of correlation between X and Y from the following data :

No. of pairs of observations n = 8, $\sum (X_t - X)^2 = 184$, $\sum (Y_i - Y)^2 = 148$, $\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 164$, X = 11 and Y=10

Using the formula, $r_{XY} = \frac{\Sigma(X-\bar{X}_i)(Y_i-\bar{Y})}{\sqrt{\Sigma(X_i-\bar{X})^2}\sqrt{\Sigma(Y_i-\bar{Y})^2}}$, we get

$$r_{xy=} \frac{164}{\sqrt{184}\times\sqrt{148}} = 0.99$$

ILLUSTRATION 3

Calculate the coefficient of correlation between age group and rate of mortality from the following data :

Age group	:	0-20	20-40	40-60	60-80	80-100
Rate of Mortality	:	350	280	540	760	900

SOLUTION

Since class intervals are given for age, their mid-values shall be used for the calculation of r.

TABLE FOR CALCULATION OF

Age group	M.V. (X)	Rate of Mor t.(Y)	$v = \frac{X_i - 50}{20}$	$v = \frac{Y_i - 50}{20}$	$u_i v_i$	$u^2 i$	$v^2 i$
0-20	10	350	–2	–19	38	4	361
20-40	30	280	–1	–26	26	1	676
40-60	50	540	0	0	0	0	0
60-80	70	760	1	22	22	1	484
80-100	90	900	2	36	72	4	1296
Total			0	13	158	10	2817

Here n=5. Using the formula (10) for correlation, we get

$$r_{xy} = \frac{5 \times 158 - 0 \times 13}{\sqrt{5 \times 10 - 0^2}\ \sqrt{5 \times 2817 - 13^2}} = 0.95$$

SHORT-CUT METHOD

When coefficient of correlation is calculated by direct method, the deviations in the X and Y series are obtained from the actual arithmetic mean of the concerned series. When arithmetic mean is not a whole number, or series are large, the deviations may be found from the assumed arithmetic means to avoid complex calculations. Later on correlations are made in $\Sigma di\, dy$, $\Sigma d^2 x$ and Σd^2_y. In calculating coefficient of correlation by short-cut method, the following formula is used:

$$r = \frac{\Sigma d_x - \frac{\Sigma d_x \times \Sigma dv}{N}}{\sqrt{\left[\Sigma d^2{}_x - \frac{(\Sigma d_x)^2}{N}\right] + \left[\Sigma dv^2 \frac{(\Sigma d_x)^2}{N}\right]}} \quad \ldots\ldots \quad (1)$$

$$r = \frac{N \Sigma d_x d_y - (\Sigma d_x) \times (\Sigma d_y)}{\sqrt{[N.\Sigma d_x{}^2 - (\Sigma d_x)^2] \times [N.\ \Sigma dy^2 - (\Sigma d_y)^2]}} \quad \ldots\ldots\ldots \quad (2)$$

In these formulae:

$\sum dx$ and $\sum dy$ = Sum of the deviations from the assumed mean in X and Y respectively.

$\sum d^2_x$ and $\sum d^2_y$ = Sum of the squares of the deviations from the assumed mean in X and Y series respectively.

Illustration 4 : Calculate the coefficient of correlation for the following age of husbands and wives :

Husbands' age	:	23	27	28	29	30	31	33	35	36	39
Wives' age	:	18	22	23	24	25	26	28	29	30	32

Husbands' age	Deviations from assumed mean	Deviations Squared	Wive's age	Deviations from assumed mean	Deviations Squared	Product of deviations
X	A_X =30 (X-30)=d_X			A_{ij} = 25 (Y-25)d_{ij}		
23	–7	49	18	–7	49	+49
27	–3	9	22	–3	9	+9
28	–2	4	23	–2	4	+4
29	–1	1	24	–1	1	+1
30	0	0	25	0	0	0
31	+1	1	26	+1	1	+1
33	+3	9	28	+3	9	+9
35	+5	25	29	+4	16	+20
36	+6	36	30	+5	25	+30
39	+9	81	32	+7	49	+63
Total	+11	215		+7	163	186
N=10						

Coefficient of correlation by the application of Ist formula:

$$\frac{\Sigma d_x d_v - \frac{\Sigma d_x \times \Sigma d_y}{N}}{\sqrt{\left(\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}\right)\left(\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}\right)}}$$

$$\frac{186 - \frac{11 \times 7}{10}}{\sqrt{\left(215 - \frac{11^2}{10}\right)\left(163 - \frac{7^2}{10}\right)}} = \frac{178.3}{\sqrt{(215\text{-}12.1) \times (163\text{-}4.9)}}$$

$$= \frac{178.3}{\sqrt{202.9 \times 158.1}} = \frac{178.3}{\sqrt{32078.49}} = \frac{178.3}{179.1} = .99 \text{ approximately}$$

Coefficient of correlation by the application of 2nd formula

$$= \frac{N.\Sigma d_x d_y - (\Sigma d_x) - (\Sigma d_x) \times (\Sigma d_y)}{\sqrt{[N.\Sigma d^2_x - (\Sigma d_x)^2] \times (N.\Sigma d^2_y - (\Sigma d_y)^2]}}$$

$$= \frac{10 \times 186 - (11 \times 7)}{\sqrt{[10 \times 215\text{-}11^2] \times [10 \times 163\text{-}7^2]}}$$

$$= \frac{1783}{\sqrt{2029 \times 1581}} = \frac{1783}{\sqrt{3207849}} = \frac{178.3}{1791} = +.99 \text{ approximately}$$

Illustration 5 : The following table gives the distribution of the total population and those who are wholly or partialy blind among them. Find out if there is any relationship between age and blindness.

Age :	0-10	10-20	2030	30-40	40-50	50-60	60-70	70-80
No. of Persons								
(in Thousand) :	100	60	40	36	24	11	6	5

Blinds :	55	40	4	040	36	22	18	15

Solution:

Note : In order to make the data comparable it is necessary to find out the number of blinds out of a fixed number (a common unit). Here this unit will be one lakh.

Age X series	Mid-value X	$A_x=45$ $(X-A_x)$ $= d_x$	d^2_x	No. of Person in thousand	No. of blind	Blind per Lakh Y	$A_y=150$ $(X-A_y)$ $= dy$	dy^2	d_xd_y
0-10	5	-40	1600	100	55	55	-95	9025	3800
10-20	15	-30	900	60	40	67	-83	6889	2490
20-30	25	-20	40	40	40	100	-50	2500	1000
30-40	35	-10	100	36	40	111	-39	1521	390
40-50	45	0	0	24	36	150	0	0	0
50-60	55	10	100	11	22	200	50	2500	500
60-70	65	20	400	6	18	30	150	22500	3000
70-80	76	30	900	5	15	500	350	122500	1050
	N=8	-40	4400			N=8	283	167435	21680

$$= \frac{\frac{\Sigma d_x d_y}{N} - \left(\frac{\Sigma d_x}{N}\right)\left(\frac{\Sigma d_x}{N}\right)}{\sqrt{\left[\frac{\Sigma d_x^2}{N} - \left(\frac{(\Sigma d_x)^2}{N}\right)\right] \times \left(\frac{(\Sigma d_y^2}{N} - \left(\frac{\Sigma d_y}{N}\right)^2\right]}}$$

$$= \frac{\frac{21860}{8} - \left(\frac{-40}{8}\right)\left(\frac{283}{8}\right)}{\sqrt{\left[\frac{4400}{8}\left(\frac{-40}{8}\right)\right]\left(\frac{167435}{8}\left(\frac{283}{8}\right)\right]}}$$

$$= \frac{288.725}{(550-25) \times (20929.375 - 1251.39}$$

$$= \frac{288.725}{/103309.39} = \frac{288.725}{3214.17} = .898 \text{ approx.}$$

Illustration 6 : Calculate coefficient of correlation of the following by the product moment method :

X	:	8	6	4	3	4
Y	:	9	7	4	4	6

SOLUTION :

	X	Y	X^2	Y^2	XY
	8	9	64	81	72
	6	7	36	40	42
	4	4	16	16	16
	3	4	9	16	12
	4	6	16	36	24
Total	25	30	141	198	166
Average	5	6	28.2	39.6	33.2

$$r = \frac{\Sigma XY - \frac{\Sigma X.\Sigma Y}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right] \times \left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}$$

$$r = \frac{166 - \frac{25 \times 30}{5}}{\sqrt{\left(141 - \frac{25^2}{5}\right)\left(198 - \frac{30^2}{5}\right)}}$$

$$= \frac{166 - 150}{\sqrt{(141-125) \times (198-180)}}$$

$$= \frac{16}{\sqrt{16 \times 18}} = \frac{16}{\sqrt{288}} = \frac{16}{16.97} = +.94 \text{ approximately}$$

2nd Formula :

$$r = \frac{N\sum XY - (\sum X) - (\sum Y)}{/[N\sum X^2 - (\sum X)^2] \times (N\sum Y^2 - (\sum d)^2]}$$

$$= \frac{5\times166 - 25 \times 30}{/[5\times141\text{-}25^2] \times [5\times198\text{-}30^2]}$$

$$\frac{830-750}{/(705 - 625) \times (990\text{-}900)} = \frac{80}{/80 \times 90} = \frac{80}{/7200}$$

$$\frac{80}{/84.853} = +.94$$

CORRELATION IN GROUPED SERIES

If the values of two variables are grouped and the frequencies of different groups are given, Karl Pearson's coefficient of correlation can be calculated of such data. This kind of classified series is called correlation table or bivariate frequency table. Such a need arises when the number of observation is very large and the data have been classified according to two measurements in a bivariate frequency table to make them concise. In such a frequency table, each cell frequency refers to both the variables. An example of such a table is given on the next page:

X	Number of Workers per Age group (Years)					
Y	20-30	30-40	40-50	50-60	60-70	Total
Days lost						
1-2	4	–	4	–	–	8
3-4	8	8	4	4	–	24
5-6	4	12	16	–	–	32

7-8	–	4	8	4	–	16
9-10	–	–	12	–	–	12
11-12	–	–	–	7	1	8
Total	16	24	44	15	1	100

The total frequency in the above table is 100. Each cell frequency is related to both X and Y variables. For example, the number of workers whose ages are between 20-30 years and have lost between 1-2 days is 4. Similarly, other frequencies are related to X-class given above and Y-class given on left.

$$r = \frac{\Sigma fd_x d_y - \frac{\Sigma fd_x \Sigma fd_y}{N}}{\sqrt{\left[\Sigma fd_x^2 - \frac{(\Sigma fd_x)^2}{N}\right] \times \left[\Sigma fd_y^2 - \frac{(\Sigma fd_y)^2}{N}\right]}}$$

$$r = \frac{N\Sigma fd_x d_y - \Sigma fd_x \times \Sigma fd_y}{/[N \times \Sigma fd_x^2 - (\Sigma fd_x)^2] \times [N \times \Sigma fd^2_y - (\Sigma fd_y)^2}$$

If step-deviations are taken, then it is not necessary to multiply the common factor (i), because both denomerator and numerator have to be multiplied by ($i_X \times i_V$), hence the value will remain the same.

Illustration 7 : Calculate the Coefficient of correlation between the age of husbands and wives from the undernoted data and comment upon the result so obtained.

Ages of Husbands	Ages of Wives					
	10-20	20-30	30-40	40-50	50-60	Total
10-20	6	3	–	–	–	9
20-30	3	16	10	–	–	29
30-40	–	10	15	7	–	32

40-50	–	–	7	10	4	21
50-60	–	–	–	4	5	9
Total	9	2	32	21	9	100

SOLUTION:

$$r = \frac{\Sigma fd_x d_y - N\left(\frac{\Sigma fd_x}{N}\right)\left(\frac{\Sigma fd_y}{N}\right)}{\sqrt{\left[\frac{\Sigma fd_x^2}{N} - \left(\frac{\Sigma fd_x}{N}\right)^2\right] \times \left(\frac{\Sigma fd_y^2}{N} - \left(\frac{\Sigma fd_y}{N}\right)^2\right]}}$$

Solution :

Y / X			Class	10-20	20-30	30-40	40-50	50-60		
			Mid point (m)	15	25	35	45	55		
			d_y	-20	-10	0	+10	+20		
Class	Mid-point (m)	d_x	d_x i=10	-2	-1	0	+1	+2	Total	fd_x
10-20	165	–20	–2	+4 6 +24	+2 3 +6	–	–	–	9	-18
20-30	25	–10	–1 +6	+2 3 +16	+1 16 0	0 10	–	–	29	–29
30-40	35	0	0	–	0 10 0	0 15 0	0 7 0	–	32	0
40-50	45	+10	+1 0	– +10	– +8	0 7	+1 10	+2 4	21	+21
50-60	55	+20	+2	– +8	– +20	–	+2 4	+4 5	9	+18
		Total		9	29	32	21	9	100	-8
		fd_y		-18	-29	0	+21	+18	-8	
		fd^2_y		36	29	0	21	36	122	
		fd_xd_y		30	22	0	18	28	98	

$$r = \frac{98 - 100 \quad \left(\frac{-8}{100}\right) \quad \left(\frac{08}{100}\right)}{100\sqrt{\left[\left(\frac{122}{100}\right) - \left(\frac{08}{100}\right)^2\right] \times \left[\left(\frac{122}{100}\right) - \left(\frac{08}{100}\right)^2\right]}}$$

$$r = \frac{97.36}{100\ 1.2136 \times 1.2136} = \frac{97.36}{121.36} = +.802$$

APPLICATIONS OF KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson's coefficient of correlation is based on three assumptions:

(1) **Normality:** The correlated variables are affected by a large number of independent factors so that they acquire normality. Variables like, age, height, weight, price, supply, demand etc., are effected by such forces that a normal distribution is formed.

(2) **Causal Relationship:** There is cause and effect relationship between the forces affecting distribution of the items in the two series. If there is no such relationship, the correlation is meaningless.

(3) **Linear Nature:** It is assumed that there is a linear relationship between the variables. In other words, if the pairs of items of both the variables are plotted on a graph paper, the plotted points will form a line.

MERITS AND LIMITATIONS OF KARL PEARSON'S COEFFICIENT

Karl Pearson's method of finding correlation between two variables is the most prevalent method, because it gives a preceise and summary quantitative figure which can be meaningfully interpreted. Pearsonian coefficient of correlation gives direction (whether positive or negative) as well as degree (high, moderate or low) of the relationship between two variables. However, there are certain limitations also of this measure, such as:

1. The coefficient of correlation assumes that there is a linear relationship between the variables under study, whether such relationship exists there or not.

2. The value of the coefficient is unduly affected by extreme items.

3. The yardstick (the r lies between $\pm$ 1) need a very careful interpretation.

4. The calculation process is time consuming.

PROBABLE ERROR OF *R*

It is an old measure to test the significance of a particular value of r without the knowledge of test of hypothesis. Probable error of r, denoted by P.E. (r) is 0.6745 times its standard error. The values of 0.6745 is obtained from the fact that in a normal distribution r $\pm$ 0.6745×S.E. covers 50% of the total distribution.

According to Horace Secrist "The probable error of correlation coefficient is an amount which is added to and substracted from the mean correlation, gives limits within which the chances are even that a coefficient of correlation from a series selected at random will fall".

$$\text{Since standard error of r, i.e., } S.E._r = \frac{1-r^2}{\sqrt{n}}, \therefore P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$$

Uses of P.E. (r)

(i) It can be used to specify the limits of population correlation coefficient p (rho) which are defined as -P.E.(r) $\leq$ P.E. (r), where p denotes correlation coefficient in population and r denotes correlation coefficient in sample.

(ii) It can be used to test the significance of an observed value of r without the knowledge of test of hypothesis. By convention, the rules are :

(a) If |r| < 6 P.E. (r), then correlation is not significant and this may be treated as a situation of no correlation between the two variables.

(b) If |r| > 6 P.E. (r), then correlation is significant and this implies presence of a strong correlation between the two variables.

(c) If correlation coefficient is greater than 0.3 and probable error is relatively small, the correlation coefficient should be considered as significant.

Illustration 8 : Find out the correlation between age and playing habit from the following information and also its probable error.

Age	:	15	16	17	18	19	20
No. of students	:	250	200	150	120	100	80
Regular Players	:	20	150	90	48	30	12

Solution :

Let X_p denote age, p the number of regular players and q the number of students. playing habit, denoted by Y, is measured as a percentage of regular Players in an age group, i.e. Y= (p/q)100.

TABLE FOR CALCULATION OF R

X	q	p	Y	d_x=X-17	d_y=Y-40	d_xd_y	Σd^2_x	d^2_y
15	250	200	80	-2	40	-80	4	1600
16	200	150	75	-1	35	-35	1	1225
17	150	90	60	0	20	0	0	400
18	120	48	40	1	0	0	1	0
19	100	30	30	2	-10	-20	4	100
20	80	12	15	3	-25	-75	9	625
Total				3	60	-210	19	3950

$$r_{XY} = \frac{-6 \times 210 - 3 \times 60}{\sqrt{6 \times 19 - 9}\ \sqrt{6 \times 3950 - 3600}} = -0.99$$

$$\text{Probable error fo r, i.e. P.E. (r)} = 0.6745 \times \frac{[1-(0.99)^2]}{\sqrt{6}} = 0.0055$$

14.1.1 SPEARMAN'S RANK DIFFERENCE METHOD

The method of ascertaining the coefficient of correlation by ranks was devised by Prof. Charles Spearman. Hence it is named after him. Spearman's Rank correlation is denoted by p(rho). It is based on the ranks of the variables. Variables are assigned ranks according to their sizes. This method of finding correlation is used where:

1. It is not possible to measure the characteristic of the items, but they can be arrayed or ranked according to some attribute. For example, in beauty contest, cooking contest, flower-show contest and in similar things, the contesting items are ranked, because it is not possible to measure their attributes in quantitative terms.

2. Data are irregular or extreme items are erratic or inaccurate, because rank-correlation coefficient is not based on the assumption of normality of data.

This method is applicable only to individual observations, rather than frequency distribution. Under ranking method original values are not taken into account, therefore, the result obtained is only approximate. The formula for coefficient of correlation by rank difference method is:

$$1-\frac{6\Sigma d^2}{N(N^2-1)} \text{ or } \quad 1-\frac{6\Sigma d^2}{N^3-1}$$

Where 'p' stands for rank coefficient of correlation.

d=difference between the ranks of paired items of X and Y variables.

Σd^2 = total of squares of rank differences N = Number of pairs of items

Method of Ranking: In this method the biggest item gets the first rank, the next to it gets second rank and so on. But difficulty may be encountered where two or more items are of equal value. In such a case one of the following methods, preferably the second, should be used.

(1) **The Bracket Rank Method:** Under this method all such items having equal values are assigned in the absence of such tie, e.g,

Items	30	32	35	35	40	42
Rank	6	5	3	3	2	1

(2) **The Average Rank Method:** Under this method all items with ties are assigned the average of ranks assignable to all of them. The next item is assigned the usual rank, e.g.,

Items	30	32	35	35	40	42
Rank	6	5	3.5	3.5*	2	1

Though the 'p' calculated by ranking method varies between $\pm$ 1, yet the values of the Karl Pearson's coefficient of correlation is not usually the same as that of the coefficient of correlation as ascertained by the ranking method.

Illustration 9: Two judges in a baby-competition rank the 12 entries as follows:

Entry :	A	B	C	D	E	F	G	H	I	J	K	L
X Judge :	1	2	3	4	5	6	7	8	9	10	11	12
Y Judge :	12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between the judges?

Solution:

Entry	Rank by X	Rank by Y	$(X_R-Y_R) = d$	$(X_R-Y_R)^3 = d^2$
A	2	12	-11	121
B	2	9	-7	49
C	3	6	-3	9
D	4	10	-6	36
E	5	3	+2	4
F	6	5	+1	1
G	7	4	+3	9
H	8	7	+1	1
I	9	8	+1	1
J	10	2	+8	64
K	11	11	0	0
N=12				$\Sigma d^2=416$

$$p = 1 - \frac{6\Sigma d^2}{N^3-N} = 1 - \frac{6\times 416}{12^3-12} = 1 - \frac{2496}{1716} = -\frac{780}{1716} = -.4545$$

It indicates that the judges have fairly strong divergent likes and dislikes so far as ranking of the babies is concerned.

CORRECTION FOR TIED RANKS

If two or more items are of equal value, they are assigned average rank. An adjustment is required for each group of tied ranks. The formula for calculating rank coefficient of correlation in case of tied ranks is:

$$p = 1 - \frac{6\,[\sum d^2 + ÉÓÔ\,(m^3 - m)}{N\,(N^2-1)}$$

m represents the number of items having tied ranks.

Illustration 10 : Calculate the rank coefficient of correlation of the following data :

X	:	80	78	75	75	68	67	60	59
Y	:	12	13	14	14	14	16	15	17

Solution

X	Rank X R	Y	Rank Y R	Rank Difference $(X_R - Y_R) = d$	d^2
80	1	12	8	- 7	49
78	2	13	7	- 5	25
75	3.5	14	5	-1.5	2.25
75	3.5	14	5	-1.5	2.25
68	5	14	5	0	0
67	6	16	2	+ 4	16
60	7	15	3	+ 4	16
59	8	17	1	+ 7	49
N=8					$\Sigma d^2 = 159.50$

$$p=1-\frac{6[\Sigma d^2+\text{ÉÓÔ}(m^3-m)]}{N(N^2-1)}=1-\frac{6[159.50+\text{ÉÓÔ}(3^3-3)}{8(8^2-1)}$$

$$=1-\frac{6[159.50+.5+2}{8(64-1)}=1-\frac{6\times162}{8\times63}=1-\frac{972}{504}=1-1.928=-.928$$

Illustration 11 : Ten entries are submitted for a competition. Three judges study each entry and then list the ten in rank order. Their ranking are as follows:

Entry No	1	2	3	4	5	6	7	8	9	10
Judge A :	9	3	7	5	1	6	2	4	10	8
Ranks given by Judge B :	9	1	10	4	3	8	5	2	7	6
Judge C:	6	3	8	7	2	4	1	5	9	10

Calculate the appropriate ranks correlations to help you answer the following question:

(a) Which pair of judges agrees the most?

(b) Which pair of judges disagree the most?

SOLUTION :

Entry No.	Bank By	Rank by	Rank by	d (A&B)	d^2	d (A&C)	d^2	d (B&C)	d^2
1	9	9	6	0	0	+3	9	+3	9
2	3	1	3	+2	4	0	0	-2	4
3	7	10	8	-3	9	-1	1	+2	4
4	5	4	7	+1	1	-2	4	-3	9
5		3	2	-2	4	-1	1	+1	1
6	6	8	4	-2	4	+2	4	+4	16
7	2	5	1	-3	9	+1	1	+4	16

8	4	2	5	+2	4	-1	1	-3	9
9	10	7	9	+3	9	+1	1	-2	4
10	8	6	10	+2	4	-2	4	-4	16
N=10					48		26		88

p (A&B) =1-	$6\sum d^2$	=1–	6×48	=1–	288	= 1-.29 = +.71
	$N(N^2-1)$		$10(10^2-1)$		990	
p(A&C) = 1-	$6\sum d^2$	=1-	6×26	= 1-	156	=1-.1575 = +.8425
	$N(N^2-1)$		$10(10^2-1)$		990	
p(B&C) = 1-	$6\sum d^2$	=1-	6×88	= 1-	528	=1-.53 = + .47
	$N(N^2-1)$		$10(10^2-1)$		990	

1.5 Check Your Progress

1. Correlation analysis is a.....

2. When the values of two variables move in the opposite directions, correlation is said to be.....

3. Non-linear correlation is also called.....

4. Scatter diagram is also called.....

5. Coefficient of correlation lies between……...

1.6 SUMMARY

Correlation studies the inter-dependence or relationship between variables. It is a statistical technique that measures the nature and extent of relationship between two or more variables.

On the basis of nature of relationship between the variables, i.e, the direction in which changes take place in them or the ratio by which they change, correlation may be : Positive or Negative, Simple, Partial or Multiple, Linear or Nonlinear

The different methods of findings out correlation are : (i) Scatter Diagram, (ii) Karl Pearson's Coefficient of Correlation, (iii) Spearman's Rank Coefficient of Correlation, (iv) Coefficient of Concurrent Deviations, (v) Least Square method.

1.7 KEYWORDS

Coefficient of correlation: A statistical measure of the degree of association between two variables.

Coefficient of determination: A statistical measure of the proportion of the variation in the dependent variable that is explained by independent variable, that is, the ratio of the explained variance to the total variance.

Scatter diagram: A graph of pairs of values of two variables that is plotted to indicate a visual display of the patter of their relationship.

1.8 SELF ASSESSMENT QUESTIONS

1. Explain the meaning and significance of the concept of correlation. How would we calculate it from the statistical point of view?
2. What is Spearman's rank correlation coefficient? Bring out its usefulness.
3. Explain the various properties of correlation coefficient.
4. What is scatter diagram? How do you interpret a scatter diagram?
5. Draw a scatter diagram to represent the following data

X:	15	18	30	27	25	23	30
Y:	7	10	17	16	12	13	9

6. Calculate the coefficient of correlation between X and Y for the above data. (9.5)

[r=+.63]

7. Draw a scatter diagram of the following data and indicate whether the correlation between the variables is positive or negative.

Height (inches) :	62	72	70	60	67	70	64	65	60	70
Weight (lbs.) :	50	65	63	52	56	60	59	58	54	65

8. Calculate the Karl Pearson's coefficient of correlation r between expenditure on advertising (Rs '000) and sales (Rs lakh), from the data given below :

Advertising: Expenses	39	65	62	90	82	75	25	98	36	78
Sales :	47	53	58	86	62	68	60	91	51	84

9. Two judges, in a beauty contest, ranked the 15 participants as follows :

X:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y:	5	11	15	6	1	10	7	12	4	13	3	2	8	14	9

What degree of agreement is there between the two judges?

10. Ten competitors in a beauty contest are ranked by three judges in the following manner :

Ist Judge :	1	5	4	8	9	6	10	7	3	2
2nd Judge :	4	8	7	6	5	9	10	3	2	1
3rd Judge :	6	7	8	1	5	10	9	2	3	4

Use the rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty?

1.9 ANSWERS TO CHECK YOUR PROGRESS

1. Bivariate and Multivariate Analysis

2. Negative Correlation

3. Curvy linear correlation

4. DOT CHART

5. -1 TO +1

1.10 REFERENCES/SUGGESTED READINGS

1. Hood, R.P. : Statistics for Business and Economics, MacMillan Business Books.
2. Amir D. Aczel and Jayavel Sounderpandian, Business Statistics: Tata McGraw Hill, 5th Edition.
3. Gupta, S.P. and Gupta M.P. : Business Statistics, Sultan Chand and Sons.
4. Gupta, S.C. : Statistical Method, Himalaya Publishing House, Delhi.
5. Bhardwaj, R.S. : Business Statistics, Excel Books.

Lesson No. 2	**Author:** Dr. Pardeep Gupta
Updated By: Ms. Chand Kiran	**Vetter:** Prof. B. S. Bodla

REGRESSION ANALYSIS

STRUCTURE

2.0 LEARNING OBJECTIVES

After reading this lesson, you must be able to

1. Understand the concept of regression analysis;
2. Determine regression equations using simple linear regression by applying least square method;
3. Calculate standard error of estimate; and
4. Apply non-linear regression equations in business forecasting.

2.1 INTRODUCTION

In correlation analysis, the emphasis is on finding whether the variables move in the same direction or in the opposite direction, and the extent of association between the two variables under study. In regression analysis, we study the pattern of relationship and the closeness of the relationship in absolute terms.

In other words, the regression analysis is a statistical method to deal with the formulation of mathematical models depicting relationships amongst variables. These modeled relationships are used for the purpose of prediction and other statistical purposes. The study of relationships among variables is prevalent in many business activities. Relationship among variables manifests itself in numerous forms and in varying intensity in different fields of study. In view of this diversity of the nature of possible relationships among variables, the relationships are classified into two broad classes, namely, deterministic and non-deterministic.

The relationship among the variables that is governed by some physical law, which is express able in the form of a mathematical function is called the deterministic relationship. There is some universally recognized theoretical basis that justifies the functional form, and any divergence of the observation from this functional relationship is considered as a result of variations in experimental conditions or merely as errors of observations.

In contrast to the deterministic relationships, we have many problems where variables are found to be associated, interdependent or one variable dependent on a number of independent variables, but such dependence or inter-relationships are not governed by any well-defined physical laws.

Suppose a factory manufactures items in batches and the production manager is interested in studying the relationship between the costs of production (Y) of a batch with the batch size (X) of the production run. In any production process, it is well known that a certain component of the production cost is fixed regardless of batch size, the overhead costs and some other administrative costs belong to this

category and the variable cost which is directly proportional to the number of units produced including raw material and labour costs. Assuming the absence of any other costs, we can develop a deterministic cost-size relationship of the form $Y = F + bx$

where F is the fixed cost and b is the unit variable cost.

It should be noted that unforeseen costs such as cost due to machine breakdowns or cost due to inferior raw material, etc., are not taken into this model, and these factors make the model non-deterministic.

Suppose a researcher is interested to study the yield of paddy (Y) per acre in relation to the varying dosage (X) of, say, fertilizers, keeping other factors of production such as irrigation, soil, etc. constant. Within a specified range of variation of X, the yield of paddy (Y) will vary, but one cannot develop a precise mathematical model to study their relationships.

The few examples of empirical relationships given above are simple in nature in the sense that only two variables have been considered at a time and the interest is to study how one changes in relation to the other. There are many business problems where several variables may be interdependent or one variable of major interest may depend upon a large number of factors so that a study of relationship may require observations on a large number of variables. These problems are handled by multiple regression analysis.

2.1.1 SIMPLE LINEAR REGRESSION

A simple linear regression involves an attempt to develop a straight line or linear mathematical model to describe the relationship between two variables. The purpose of a regression equation is to estimate the values of one variable based on the known values of the other. Another use of regression equation is to predict the values of the one variable given the values of other variables. For instance, a trade unionist may attempt to explain the changes in the level of unemployment in terms of extent of automation in Indian industry. It should be noted that the logic of

causal relationship must come from theories outside the domain of statistics. Regression analysis merely indicates what, if any, mathematical relationship there might be.

Although there could be a wide variety of forms of relationships, we shall restrict our discussion, in this section, to linear equations only. Linear or straight line equations are important because they closely approximate many real world relationships and they are relatively easy to work with and interpret. However, we shall extend our discussion to other forms of regression analysis in later sections of this chapter. Let us consider a simple situation where we are interested with the relation between two variables X and Y, where X is the independent variable or the causal variable on which Y depends. Suppose we have collected a series of values of X and the corresponding values of Y. Let the series of values of X be denoted by $X_1, X_2 \ \ldots \ X_n$, and the corresponding values of variable Y be denoted by $Y_1, Y_2, \ldots Y_n$. The variables Y, apart from having a relation with X, are also affected by chance variation. These are regarded as random variables, while the chosen values of X are regarded as fixed constants.

In order to illustrate the procedure, we shall consider the problem of the production manager, i.e., to study the relation between batch size

(X) and the production cost (Y).

Table 1 gives a set of 10 values for X and Y, when only two variables are involved, the logical first step is to plot the data as points on a graph paper, or to obtain a scatter diagram.

Table 1: Data relating to batch size (X) and production costs (Y)

(in thousand Rs.)

Batch No.	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
X	11	13	18	24	28	32	38	42	47	53

Y	2.1	2.7	2.9	2.9	3.1	3.0	3.3	3.7	4.0	4.4

The scatter diagram (see Fig. 1) gives some idea of the nature of relationship to be considered, i.e., whether a straight line or a curve, and also provides a visual impression of the extent of variation about the line or the curve. The plotted data will also indicate whether the hypothetical model is likely to be confirmed, and if not, the nature of modification required to the hypothetical model. From the scatter diagram of the data of Table 1, one can say that an approximate linear model can be fitted to the data. After confirming from the scatter diagram that the relationship is approximately linear in nature, we start with developing a mathematical model and proceed to the estimation of the underlying relation. We assume that a linear relationship of the form

$$Y = a + bx_i + U_t,\ i = 1, 2\ldots n$$

exists between the variables Y and X, where U_t, i = 1, …, n, is the random error factor corresponding to the ith observation, and a and b are model

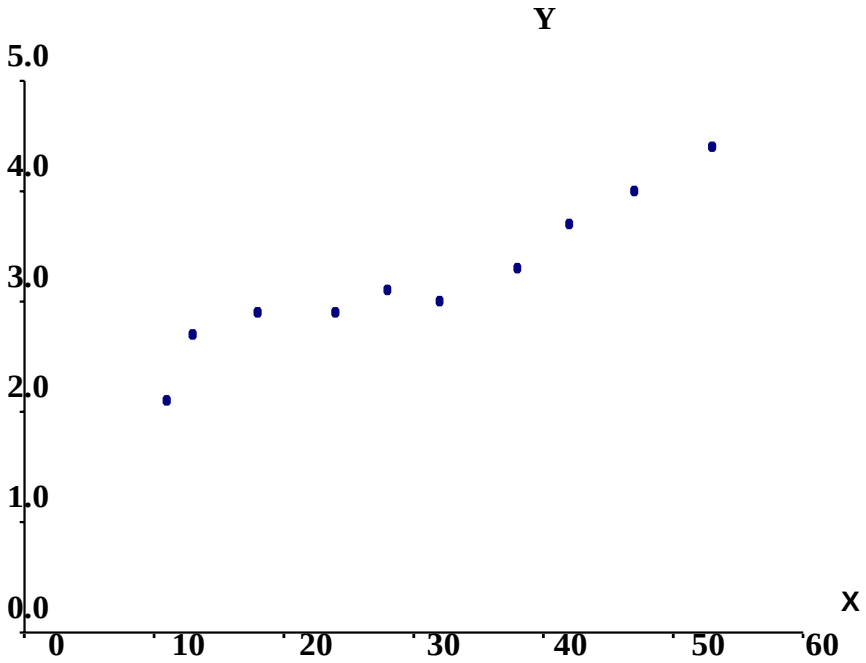


Fig. 2.1: Scatter Diagram for the Data in Table1

parameters characterizing the unknown linear relation. The error factor U_i is assumed to be independent random variable, with $E(U_t) = 0$, and Var $(U_i) = \sigma^2$ for all i = 1, 2, …, n, the variance σ^2 is unknown. According to this model, each observation Y_i is a random sample of one observation from the normal distribution with mean $= a + bx_i$, and variance σ^2.

2.1.2 LEAST SQUARES METHOD

The problem of fitting a straight line to the regression model is equivalent to estimating the regression parameters a and b from the observed data. One simple method is by the 'eye-estimation', but this method suffers from the following main drawbacks: (i) it is a subjective method, (ii) it leaves no scope for probabilistic assessment of errors, construction of confidence intervals, etc., and (iii) it cannot be used in multivariate data where a scatter diagram cannot be plotted. Even in two variable cases, this method cannot be applied if the relation between the variables appears to be non-linear. The best method that is used in estimating the regression parameters a and b is the "method of least squares", which is an objective, and efficient method of estimating regression parameters even in cases not limited to linear.

In order to estimate a and b by the method of least squares, we need to minimize

$$\sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} (Y_i - a - bx_i)^2 \qquad \ldots(1)$$

Expression (1) can also be written by

$$\sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} (Y_i - Y_c)^2 \qquad \ldots(2)$$

where Y_i = an observed value of Y, Y_c = the computed value of Y using the least squares method.

To minimize ΣU_i^2, we differentiate it partially with respect to a and b, and equate the results to zero. Thus, we have

$$\frac{\delta}{\delta a} \Sigma U_i^2 = -2 \sum_{i=1}^{n} (Y_i - a - bX_i) = 0$$

and $$\frac{\delta}{\delta b} \Sigma U_i^2 = -2 \sum_{i=1}^{n} (Y_i - a - bX_i)(X_i) = 0$$

Solving these equations, we get the "normal equations" as

$$\Sigma Y_i = na + b\Sigma X_i$$
$$\Sigma X_i Y_i = a\Sigma X_i + b\Sigma X_i^2 \qquad \ldots (3)$$

where n is the number of paired observations. Thus, by obtaining the various quantities such as ΣX_i, $\Sigma X_i Y_i$ and so on, we can solve these two simultaneous equations for a and b.

We can also find regression equation of X on Y. It can be expressed as X = a + bY. The values of a and be will be obtained from

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

Illustration 2.1. From the following data obtain the regression equation of Y on X.

X :	0	1	2	3	4
Y :	1	1.8	3.3	4.5	6.3

Solution. The regression equation of Y on X be Y = a + bX

To determine a and b, the following two normal equations are required to be solved

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

The value of $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$ can be obtained from the following table.

Table: Computations for regression coefficients

	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	6.3	25.2	16	
∑X=10	∑Y=16.9	∑XY=47.1	$\sum X^2$=30	

On substituting the values of ∑X, ∑Y, ∑XY and $\sum X^2$, we get the following normal equations

$$16.9 = 5a + 10b$$

$$47.1 = 10a + 30b$$

Solving these equations we get a = 0.72 and b = 1.33. Hence, the regression equation is Y = 0.72 + 1.33X.

Illustration 2.2. From following data obtain the two regression equations:

X :	6	2	10	4	8
Y :	9	11	5	8	7

Solution. Obtaining regression equations

X	***Y***	***XY***	***X^2***	***Y^2***
6	9	54	36	81
2	11	22	4	12 1
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
∑X=30	**∑Y=40**	**∑XY = 214**	**$\sum X^2$ = 220**	**$\sum Y^2$ = 340**

Regression equation of Y on X : $Y_C = a + bX$

To determine the values of *a* and *b* the following two normal equations are to be solved.

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

$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values $40 = 5a + 30b$ … (i)

$214 = 30a + 220b$ … (ii)

Multiplying equation (i) by 6,240 = $30a + 180b$ … (iii)

$214 = 30a + 220b$ … (iv)

Deducting equation (iv) from (iii) $- 40b = 26$ or $b = -0.65$

Substituting the value of b in equation (i)

$40 = 5a + 30 (-0.65)$ or $5a = 40 + 19.5 = 59.5$ or $a = 11.9$

Putting the values of *a* and *b* in the equation,

the regression of Y on X is $Y = 11.9 - 0.65X$

Regression line of X on Y : $X_C = a + bY$ and the

two normal equations are :

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

$$\sum XY = a\sum Y + b\sum Y^2$$

$30 = 5a + 40b$ … (i)

$214 = 40a + 340b$ … (ii)

Multiplying equation (i) by 8: $240 = 40a + 320b$ … (iii)

$214 = 40x + 340b$ … (iv)

From Eqn. (iii) and (iv)

$-20b = 26$ or $b = -1.3$

Substituting the value of *b* in equation (i); $30 = 5a + 40 (-1.3)$

$5a = 30 + 52 = 82$

$a = 16.4$

Putting the values of *a* and *b* in the equation, the regression line of X on Y is

$X = 16.4 - 1.3Y$.

- **Deviation taken from Arithmetic Mean of X and Y**

The computation for finding out regression equation of Y on X can be simplified by taking the deviation of Y and X series from their respective means. The regression equation of Y on X will, then, be as follows:

$$Y - \overline{Y} = r \frac{\sigma_y}{\sigma_x} (X - \overline{X}) \qquad \ldots (4)$$

where, $\overline{X}$ is mean of the series X, $\overline{Y}$ is mean of the series Y. The term $r\frac{\sigma_y}{\sigma_x}$ is known as regression coefficient of Y on X and is denoted as b_{yx}. It measures the change in Y corresponding to a unit charge in X. The regression coefficient b_{yx} can be calculated as follows :

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

where, x and y represent the deviations from their actual means. Therefore, for calculating the regression coefficient, instead of calculating correlation coefficient, σ_x and σ_y, we can find out its value using the above mentioned formula i.e., by calculating Σxy and Σx^2 and dividing the former by the latter.

We can also calculate the regression equation of X on Y as follows :

$$X - \overline{X} = r \frac{\sigma_x}{\sigma_y} (Y - \overline{Y}) \qquad \ldots (5)$$

The regression coefficient coefficient of X on Y, is denoted by b_{xy} and it measures the change in X corresponding to a unit change in Y.

Thus, $$b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

We can also calculate the regression coefficient of X on Y, using the following formula

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$$

where x and y are deviations from their actual means.

Illustration 2.3. From the data of illustration 2, calculate the regression equations taking deviation of items from the mean of X and Y series.

Solution. Calculation of regression equations

X	$(X - \overline{X})$ x	x^2	Y	$(X - \overline{Y})$ y	Y^2	xy
6	0	0	9	+1	1	0
2	-4	16	11	+3	9	-12
10	+4	16	5	-3	9	-12
4	-2	4	8	0	0	0
8	+2	4	7	-1	1	-2
$\Sigma X=30$	$\Sigma x=40$	$\Sigma x^2=40$	$\Sigma Y=40$	$\Sigma y=0$	$\Sigma Y^2 = 20$	$\Sigma xy=-26$

Regression Equation of X on Y : $X - \overline{X} = r \frac{\sigma_x}{\sigma_y} (Y - \overline{Y})$

$$r \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{20} = -1.3$$

$$\overline{X} = \frac{30}{5} = 6 ; \quad \overline{Y} = \frac{40}{5} = 8$$

Hence $X - 6 = -1.3 (Y-8) = -1.3 Y + 10.4$

$X = -1.3Y + 16.4$ or $X = 16.4-1.3Y$

Regression Equation of Y on X : $Y - \overline{Y} = r \frac{\sigma_x}{\sigma_y} (X - \overline{X})$

$$r \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{-26}{40} = -0.65$$

$Y - 8 = -0.65 (X-6) = -0.65X + 3.9$

$Y = -0.65X + 11.9$ or $Y = 11.9-0.65X$

Thus we find that the answer is the same as obtained earlier. However, the calculations are very much simplified without the use of normal equations.

- **Deviation taken from Assumed Means**

When actual means of *X* and *Y* variables are in fractions the calculations can be simplified by taking the deviations from the assumed means. When deviations are taken from assumed means the entire procedure of finding regression equations remains the same—the only difference is that instead of taking deviations from actual means we take the deviations from assumed means. The two regression equations are:

$$X - \overline{X} = r \frac{\sigma_x}{\sigma_y} (Y - \overline{Y})$$

The value of $r\ \sigma_x / \sigma_y$ will now be obtained as follows:

$$r \frac{\sigma_x}{\sigma_y} = \frac{N\Sigma d_x d_y - \Sigma d_x \times \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$d_x = (X - A)$ and $d_y = (Y - A)$

Similarly, the regression equation of Y on X is

$$Y - \overline{Y} = r \frac{\sigma_y}{\sigma_x} (X - \overline{X})$$

$$r \frac{\sigma_y}{\sigma_x} = \frac{N\Sigma d_x d_y - \Sigma d_x \times \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

It should be noted that in both the cases the numerator is the same, the only difference is in the denominator.

Illustration 3.4. From the data of illustration 2 calculate regression equations by taking deviations of X series from 5 and of Y series from 7.

X	$(X-5)$ d_X	d_X^2	Y	$(X-7)$ d_y	d_y^2	$d_x d_y$
6	+1	1	9	+2	4	+2
2	-3	9	11	+4	16	-12
10	+5	25	5	-2	4	-10
4	-1	1	8	+1	1	-1
8	+3	9	7	0	0	0
$\sum X=30$	$\sum d_X=+5$	$\sum d_X^2=45$	$\sum Y=40$	$\sum d_y=5$	$\sum dy^2=25$	$\sum d_x d_y=-21$

Regression Equation of X on Y : $X - \overline{X} = r \frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$

$$\overline{X} = \frac{\sum X}{N} = \frac{30}{5} \; ; \qquad \overline{Y} = \frac{40}{5} = 8$$

$$r \frac{\sigma_x}{\sigma_y} = \frac{N\sum d_x d_y - \sum d_x \times \sum d_y}{N\sum d_y^2 - (\sum d_y)^2}$$

$$= \frac{5(-21) - (5)(5)}{(5)(25) - (5)^2} = -\frac{105-25}{125-25} = -\frac{130}{100} = -1.3$$

$$X - 6 = -1.3(Y-8)$$

$$X - 6 = -1.3Y + 10.4 \quad \text{or} \quad X = 16.4 - 1.3Y$$

Regression equation of Y on X: $Y - \overline{Y} =$

$$r \frac{\sigma_x}{\sigma_y} (X - \overline{X})$$

$$r \frac{\sigma_y}{\sigma_x} = \frac{N\Sigma d_x d_y - \Sigma d_x \times \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{5(-21) - (5)(5)}{(5)(45) - (5)^2} = \frac{-105-25}{200} = -0.65$$

$$Y - 8 = -0.65 (X-6)$$

$$Y-8 = -0.65X + 3.9 \quad \text{or} \quad Y = 11.9-0.65X$$

It is clear from this example that answer would come out to be the same whether we take deviations from actual means or assumed means.

If the regression coefficient of X on Y and the regression coefficient of Y on X, are known, the correlation coefficient can be calculated by taking the underroot of the product of two regression coefficients. Thus,

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

$$\text{Since} \quad b_{xy} = \frac{r\sigma_x}{\sigma_y} \quad \text{and} \quad b_{yx} = \frac{r.\sigma_y}{\sigma_x}$$

$$\text{therefore,} \quad b_{xy} . b_{yx} = \frac{r.\sigma_x}{\sigma_y} . \frac{r.\sigma_y}{\sigma_x} = r^2$$

$$\text{Hence,} \quad r = \sqrt{b_{xy} . b_{yx}} \qquad \ldots (6)$$

The following are the important properties of the regression coefficients:

(i) Both regression coefficients will have the same sign, i.e. either both are positive or both are negative.

(ii) Both regression coefficients cannot be greater than one, because, the value of correlation (r) cannot exceed one.

(iii) The coefficient of correlation will have the same sign as that of regression

coefficients. Thus, for $b_{xy} = -0.6$ and $b_{yx} = -1.4$,

the value of r would be . $\sqrt{-0.6 \times -1.4}$

The sign of the coefficient, however, would be negative and not positive.

(iv) The regression coefficients are independent of change of origin but not of scale.

Illustration 2.5. 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 year : 2 4 6 8

Maintenance cost (Rs '00) : 10 20 25 30

Estimate the maintenance cost for a ten year old car.

Solution. Let X denote the age and Y denote the maintenance cost. The calculations are given in Table below:

X	$(X - \overline{X})$ x	x^2	Y	$(Y - \overline{Y})$ y	y^2	xy
2	–3	9	10	–11.25	126.5625	33.75
4	–1	1	20	–1.25	1.5625	1.25
6	1	1	25	3.75	14.0625	3.75
8	3	9	30	8.75	76.5625	26.25
20		20	85		218.75	65.00

$$\overline{X} = \frac{20}{4} = 5, \qquad \overline{Y} = \frac{85}{4} = 21.25$$

$$\frac{r\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{65}{20} = 3.25$$

The regression equation is

$$Y - \overline{Y} = [\{(r\sigma_y)/\sigma_x\} \cdot (X - \overline{X})$$
$$Y - 21.25 = 3.25\ (X - 5)$$
$$Y - 21.25 = 3.25\ X - 16.25$$
$$Y = 3.25\ X + 5$$

The maintenance cost for a ten year old car will be :

$$Y = 3.25 \times 10 + 5$$
$$= 37.5 \text{ or Rs. } 3750.$$

Illustration 15.6. In a correlation study the following values are obtained.

	X	***Y***
Mean	65	67
Standard deviation	2.5	3.5
Coefficient of correlation	0.8	

Find the two regression equations that are associated with the above values.

Solution. The regression equation of X on Y

$$X - \overline{X} = \frac{r\sigma_x}{\sigma_y} (Y - \overline{Y})$$

We are given, $\overline{X} = 65$, $\overline{Y} = 67$, $r = 0.8$, $\sigma_x = 2.5$, $\sigma_y = 3.5$

Substituting the values,

$$X - 65 = 0.8 \times \frac{2.5}{3.5} (Y - 67)$$

$$X = 65 + 0.5714 (Y - 67)$$

$$= 65 + 0.5714 Y - 38.28$$

$$= 0.5714 Y + 26.72$$

The regression equation of Y on X,

$$Y - \overline{Y} = \frac{r\sigma_y}{\sigma_x} (X - \overline{X})$$

i.e. $$Y - 67 = 0.8 \times \frac{3.5}{2.5} (X - 65)$$

$$Y = 67 + 1.12 (X - 65)$$

$$= 67 + 1.12 X - 72.8$$

$$= 112 X - 5.8$$

2.2 STANDARD ERROR OF ESTIMATE

The standard error of estimate indicates how precise the prediction of Y is, based on the regression equation of Y on X. The standard error of estimate denoted as S_{yx} is defined as:

$$S_{yx} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n}} \qquad \ldots (7)$$

More accurate formula for small samples and the value used in analysis of variance is given by

$$S_{yx} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n-2}} \qquad \ldots (8)$$

where Y is the actual value, Y_C is the predicted value using the regression equation. Thus, this concept is similar to the concept of standard deviation which measures the dispersion about an average such as the mean. The standard error of estimate is a measure of dispersion about the average relationship given by the regression line.

The other formula for computation of standard error of estimate are:

(i) $S_{yx} = \sigma_y\sqrt{1 - r^2}$

(ii) $S_{yx} = \sqrt{\frac{Y^2 - a\Sigma Y - b\Sigma XY}{n}}$

The standard error of estimate helps in ascertaining as to how much representative the regression equation is as a description of the average relationship between two variables. The smaller is the value of standard error of estimate, the better is the fit of the equation to the given data and better are the estimates based on the

regression equation. If the standard error of estimates is zero, there is a perfect match and hence, it is a case of perfect correlation.

Illustration 2.7. Data related to yield on Security (Y) and yield on market portfolio (X) is given as follows:

Period	***Yield on Security (Y)***	***Yield on market portfolio (X)***
1	6.2	5.4
2	7.0	6.7
3	7.2	6.8
4	7.8	8.0
5	7.0	5.02
6	7.2	5.0
8	7.2	6.2
9	7.3	6.5
10	7.1	6.1

Calculate standard error or estimate.

Solution. The normal equations are:

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + \sum X^2$$

Substituting the values of $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$,

the normal equations are

$71.4 = 10a + 61.12b$

$438.24 = 61.12a + 381.55b$

Solving these equations we get,

$a = 5.7294$ and $b = 0.2308$.

Hence, the regression equation is,

$Y = 5.7294 + 0.2308\ X$

Substituting the values of X in the above equation, we can get the values of Y. These calculated values of Y are given as follows:

Table: Estimated values of Y

Period	***Y***	Y_C	$Y - Y_C$	$(Y - Y_C)^2$
1	6.2	6.97	0.77	0.5929
2	7.0	7.27	0.27	0.0729
3	7.2	7.30	0.10	0.0100
4	7.8	7.57	0.23	0.0529
5	7.0	6.89	0.11	0.0121
6	7.2	6.88	0.32	0.1024
7	7.4	6.97	0.43	0.1849
8	7.2	7.16	0.04	0.0016
9	7.3	7.23	0.07	0.0049
10	7.1	7.14	0.04	0.0016

The standard Error of Estimate

$$= \sqrt{\frac{\Sigma(Y - Y_c)^2}{n}} = \sqrt{\frac{1.0362}{10}} = 0.3219$$

2.2.1 Coefficient of Determination

In order to measure the extent, or strength, of the association that exists between the two variables, X and Y, we use the statistic called the coefficient of determination. Since we use sample points to develop regression lines, we refer to this measure as the sample coefficient of determination. This measure is developed on the basis of two kinds of variations: (i) the variation of Y value around the fitted regression line and (ii) the variation of the Y values around their

own mean. The sum of a group of squared deviations is termed as the variation here. Therefore, the variation of the Y values around the regression line is given by

$$\sum(Y-\hat{Y})^2$$

and the second variation, that is, the variation of Y values around their own mean is

$$\sum(Y-\overline{Y})^2$$

The sample coefficient of determination is, therefore, given by

$$R^2 = 1 - \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\overline{Y})^2} \qquad \ldots (9)$$

R^2 gives the proportion of variation in Y that can be accounted for by the variation in X. The value of R^2 can lie between 0 and 1. Whenever the regression line is a perfect estimator, the value of R^2 is equal to 1. It should be noted that an R^2 close to 1 indicates a strong correlation between X and Y, while an R^2 near 0 means there is little correlation between the two variables.

Another way of interpreting the sample coefficient of determination is by considering the amount of the variation in Y that is explained by the regression line. The total variation, that is, the sum of the square total deviations of the observed points from their mean would be

$$\sum(Y-\overline{Y})^2$$

and the explained portion of the total variation, or the sum of the square of explained deviations of these points from their mean, would be

$$\sum(Y-\overline{Y})^2$$

Therefore, the fraction of the total variation that remains unexplained would, then, be

$$\frac{\sum(\mathbf{Y}-\hat{\mathbf{Y}})^{\mathbf{2}}}{\sum(Y-\overline{Y})^2}$$

Subtracting the fraction of the total variation that remains unexplained from one, we will get the formula for finding that fraction of the total variation of Y which is explained by the regression line Y as same as Y_C. Thus, we have

$$R^2 = 1 - \frac{\Sigma(Y-\hat{Y})^2}{\Sigma(Y-\overline{Y})^2} \qquad \ldots (10)$$

We can state here that R^2 measures how well X explains Y. A simplified formula for the sample coefficient of determination is given by

$$R^2 = \frac{a\Sigma Y + b\Sigma XY - n\overline{Y}^2}{\Sigma(Y^2-\overline{Y})^2}$$

where a and b are regression coefficients, n is the number of sets of observations.

2.3 NON-LINEAR REGRESSION AND LINEARIZATION

So far we have studied situations where the underlying relationship between a dependent variable and an independent variable X could be approximately formulated in terms of a simple linear regression model. However, very often when we plot the scatter diagram of Y versus X, it indicates that the relationship cannot be exactly approximated by a straight line, and hence the linear regression analysis cannot be directly used. If the form of the non-linear relationship between two variables is suggested by some theoretical statistical approach, it is sometimes possible to make a transformation of one or both the variables such that the relationship between the transformed variables can be expressed as a straight line. For example, if we know that the relationship between X and Y is of the form

$$Y = ab^X \qquad \ldots (11)$$

then, taking logarithms of both sides of (11), we get

$$\log Y = \log a + X \log b \quad \ldots (12)$$

or $Y` = A + BX$

where, $Y` = \log Y$, $A = \log a$ and $B = \log b$

The relationship between Y and X, as expressed by (12), is linear and hence with the usual approach of a linear model, regression analysis can be extended to the transformed variables.

Illustration 2.8. Form a regression equation $Y = ab^x$ from the following data:

x :	2	3	4	5	6
y :	144	172.8	207.4	248.8	298.5

Solution. Taking logarithm of $Y = ab^x$, we get $\log Y = \log a + X \log b$

The normal equations are

$$\sum \log Y = n \log a + \log b \sum X$$

$$\sum X \log Y = \log a \sum X + \log b \sum X^2$$

the calculations of $\sum X$, $\sum \log Y$, $\sum X^2$ and $\sum X \log Y$ are given as follows :

Table : Computations for non-linear regression

X	***X²***	***Y***	***log Y***	***X log Y***
2	4	144.0	2.1584	4.3168
3	9	172.8	2.2375	6.7125
4	16	207.4	2.3168	9.2672
5	25	248.8	2.3959	11.9795
6	36	298.5	2.4749	14.8494
20	90		11.5835	47.1254

Therefore, the normal equations are:

$$11.5835 = 5 \log a + 20 \log b$$

$$47.1254 = 20 \log a + 90 \log b$$

Solving these equations and taking antilogarithms we get $a = 100$ and $b = 1.2$.

Therefore, the regression equation is $Y = 100 (1.2)^{x}$.

2.4 DIFFERENCE BETWEEN REGRESSION AND CORRELATION

The discussion in the preceding sections brings out the fact that the two techniques of regression and correlation are based on different set of assumptions. In practice, it is not always clear which should be used in a given situation. It can again be seen from the discussion in the preceding sections that the correlation coefficient r is related to the coefficient of determination R^2 in regression analysis. Also the correlation coefficient r takes the same sign as that of b. Thus, the accuracy of the regression equation as a prediction device depends on the degree of covariance between the two variables. However, there are several differences between correlation and regression. These are given as follows:

(i) The regression analysis is mainly aimed at establishing the functional relationship between the dependent and independent variables such that the relationship can be used for prediction of dependent variable on the basis of independent variable(s). Correlation aims at measuring the degrees of variation in X and Y but it does not imply a functional relationship. Therefore, correlation coefficient r does not make it possible to predict Y on the basis of X.

(ii) It may be noted that a given value of correlation coefficient r is consistent with an infinite number of regression lines. Even if intercept is changed, the value of r still remains the same. This indicates that r merely measures the strength of relationship but it does not give the equation for the regression line.

(iii) In the regression analysis, Y is assumed to be a random variable while X's

are control variables (fixed quantities). Correlation analysis is based on the assumption of joint probability distribution of X and Y when both are random variables. This implies that regression analysis is used if one variable is clearly dependent upon the other, or one is measured subsequent to the other. In correlation analysis, neither of the variables can be considered as being subsequent to or as being a consequence of the other.

In practice, the choice between the regression and the correlation analysis depends on the purpose of investigation. If the purpose is merely to determine the degree of association, the correlation may be used. On the other hand, if the purpose is to predict the level of dependent variable given the value of independent variable(s), the regression analysis is used. Correlation analysis may, however, be used as the starting point for selecting useful independent variables for regression analysis.

2.5 CHECK YOUR PROGRESS

1. In regression analysis, the variable that is being predicted is the

a. response, or dependent, variable

b. independent variable

c. intervening variable

d. is usually x

2. If the coefficient of determination is a positive value, then the regression equation

a. must have a positive slope

b. must have a negative slope

c. could have either a positive or a negative slope

d. must have a positive y intercept

3. In regression analysis, if the independent variable is measured in kilograms, the dependent variable

a. must also be in kilograms

b. must be in some unit of weight

c. cannot be in kilograms

d. can be any units

4. How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

a) 1

b) 2

c) 3

d) 4

5. In a simple linear regression model (One independent variable), If we change the input variable by unit. How much output variable will change?

a) By 1

b) No change

c) By its Slope

d) one of the above

2.6 SUMMARY

The regression analysis is a statistical method to deal with the formulation of mathematical models depicting relationships amongst variables. There modeled relationships are used for the purpose of prediction and other statistical purposes. A simple linear regression involves an attempt to develop a straight line or linear mathematical model to describe the relationship between two variables. The purpose of a regression equation is to estimate the values of one variable based on the known values of the other. The scatter diagram gives some idea of the nature of relationship to be considered i.e., whether a straight line or a curve and also provide a visual impression of

the extent of variation about the line or the curve. Least square method is considered the best method of estimating the regression parameters a and b. The linear regression equation for Y on X may be written as follows:

$$Y_C = a + bX$$

The values of a and b can be found by solving the following two equations:

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

The computations for finding out regression equation of Y on X can be simplified by taking deviation of Y and X series from their respective means. The regression equation of Y on X will then be:

$$Y - \overline{Y} = r\frac{S_y}{S_x}(X - \overline{X})$$

The term $r\frac{S_y}{S_x}$ is known as regression coefficient of Y on X and is denoted by by_x and calculated as follows:

$$b_{yx} = \frac{\sum xy}{\sum x^2}$$

The standard error of estimate indicates how precise the prediction of Y is, based on the regression equation of Y on X. The standard error is computed as follows:

$$S_{yx} = \sqrt{\frac{\Sigma(Y - Y_C)^2}{n}}$$

The statistic-coefficient of determination is used to measure the extent of association between the variables X and Y.

2.7 KEYWORDS

Dependent variable: The variable which is being predicted or explained by the independent variable.

Independent variable: The variable which can be decided independently and then used to predict or explain a dependent variable.

Least squares line: The straight line around which the sum of squared residuals is minimum.

Regression: A description of the relationship between two continuous variables.

Regression equation: The mathematical equation relating the dependent and independent continuous variables.

Standard error of estimate: A measure of variation of the observed values of the dependent variable about the least squares regression line.

2.8 SELF ASSESSMENT TEST

1. What do you understand by regression analysis? Explain its significance.

2. The following marks (out of 100) have been obtained by students in Economics and Statistics :

   Economics : 80 45 55 56 58 60 65 68 70 75 85

   Statistics : 82 56 50 48 60 62 64 65 70 74 90

   Find the lines of regression.

3. Given the bivariate data

   X : 1 5 3 2 1 1 7 3

   Y : 6 1 0 0 1 2 1 5

   (a) Fit a regression line of Y on X and hence predict Y if X = 5.

   (b) Fit a regression line of X on Y and hence predict X if Y = 2.5.

4. Is the following statement correct? Give reasons. The regression coefficient of x on y is 3.2 and that of y on x is 0.8.

5. Find the most likely price of a firm's share in Bombay stock exchange corresponding to the price of Rs. 70 at Ahmedabad stock exchange from the following :

	Ahmedabad stock exchange	***Bombay stock exchange***
Average price	65	67
Standard deviation	2.5	3.5

Coefficient of correlation between the two prices of the firm's share in the two exchanges is 0.8

2.9 ANSWERS TO CHECK YOUR PROGRESS

1. A

2. C

3. D

4. B

5. C

2.10 REFERENCES/SUGGESTED READINGS

1. Hood, R.P. : Statistics for Business and Economics, MacMillan Business Books.

2. Amir D. Aczel and Jayavel Sounderpandian, Business Statistics, Tata McGraw Hill, 5th Edition.

3. Gupta, S.P. and Gupta M.P. : Business Statistics, Sultan Chand and Sons.

4. Gupta, S.C. : Statistical Method, Himalaya Publishing House, Delhi.

5. Bhardwaj, R.S. : Business Statistics, Excel Books.

Lesson No. 3	**Author:** Dr. Karam Pal Singh
Updated By: Ms. Chand Kiran	**Vetter:** Prof. B. S. Bodla

PROBABILITY

STRUCTURE

3.0 LEARNING OBJECTIVES

AFTER READING THIS LESSON, YOU SHOULD BE ABLE TO–

- Understand concepts and approaches to probability;
- Calculate probability for dependent, independent and mutually exclusive events; and
- Discuss additive, multiplicative and Bayesian approaches to probability.

3.1 INTRODUCTION

The word probability or chance is very common in day to day life of human being. For instance, we come across statements like "India may win the World Cup"; "It is likely that Mr X may not come for teaching statistics class today"; "Probably US may win the War against Terrorism". All these terms possible, probable, likely, etc, convey the same sence i.e. the event is not certain to take place or there are some uncertainties about the happening of the events.

In simple words, the word probability thus refers that there is uncertainty about the happening of event. The theory of probability has its origin in the games of chance related to gambling such as throwing a die, tossing a coin, drawing cards from a pack of cards etc. Jerane Cardon (1501-76), an Italian mathematician, was the first scholar to write a book on the subject entitled "Book on Games of Chance" which was published after his death in 1663. During the last quarter of the eighteenth century, the study of the games of chance no longer remained dependent on the initiative of the gamblers. The subject became an area of academic interest and a number of scholars addressed themselves to the field of probability. In the early nineteenth century the famous French mathematician Laplace, and the German mathematician Gauss carried the knowledge of the subject many important9s9teps forward. With the expansion of national economics, some great persons like De Moivre (1718), James Bernoulli (1713), Bayes (1768) etc., were inspired to develop the theory of probability further and apply it in different fields of decision making. In later years,

R.A. Fisher, Karl Pearsons, J. Neyman etc. developed a sampling distribution theory based on the laws of probability.

Today a comprehensive theory of probability exists and in the words of Emile Borel, "Probability theory is of interest, not only to card and dice players, who were its godfathers, but also to all men of action, heads of industries or heads of armies, where success depends on two sorts of factors the one known or calculable, the other uncertain and probabilistic."

3.1.1 MEANING AND DEFINITION OF PROBABILITY

One of the major reasons for the evolution and development of the theory of probability is its presence in almost every aspect of practical life. A phenomenon is random if chance factors determine its outcome.

All the possible outcomes may be known in advance, but the particular outcome of a single trial in any experimental operation cannot be pre-deter- mined. Nevertheless, some regularity is built into the process so that each of the possible outcomes can be assigned a probability fraction. The simplest example of a random phenomenon is the result of the toss of a coin. Though all the possible outcomes are known, i.e., head or tail, but chance factors determine the outcome of any single toss. There is no deterministic regularity here, that is, one cannot say for sure that head or tail, will come up on a particular toss. Similarly, in the roll of a cubic die, we cannot predetermine which side will turn up; even though, all the possible outcomes are known. The existence of random phenomena is found in so many diversified fields that it is imperative particularly for students in the social sciences to study the theory of probability.

Probability is especially important in statistics because of the many principles and procedure that are based on this concept. Indeed, probability plays a special role in all our lives, and has an everyday meaning. Sometimes we hear phrases like: 'You had better take an umbrella because it is likely to rain.' 'His chances of winning are pretty small.' It is very likely that it may rain by the evening. You are probably right.' or 'There are fifty-fifty chances of his passing the examination.' In each of these phrases an idea of uncertainty is acknowledged. Goethe remarked that, "There is nothing more frightful than action in ignorance." Reasoning in terms of probabilities is one weapon by which we attempt to reduce this uncertainty or ignorance. The use of word 'probability' in statistics, however is somewhat different. It is more precise than what it means in popular usage. In statistics, a probability is a numerical value that measures the uncertainty that a particular event will occur.

3.2 APPROACHES TO PROBABILITY

There are basically four approaches for measuring the probability. They represent different conceptual aspects for understanding the gamut of probability. They are:

- Classical Approach.
- Empirical Approach.
- Subjective Probability.
- Modern Approach

Let us discuss them in detail for the sake of smooth understanding of the students.

3.2.1 CLASSICAL APPROACH TO PROBABILITY

Since the theory of probability had its origin in gambling games, the method of measuring probabilities, which was just developed, was particularly appropriate for gambling situations. This method of measuring probability is called classical or a priori concept of probability. In such situations, the probability of an event is simply the ratio of number of favourable out- comes of an event to the number of possible outcomes, where each outcome is equally likely to occur. In other words, if there are several equally likely events that may happen, the probability that any one of these events will happen, is the ratio of the number of cases fovourable to its happening to the total number of possible cases. If there are 'n' possible outcomes favourable to the occurrence of an event 'A' and 'm' possible outcomes unfavorable to the occurrence of A, and all of these possible outcomes are equally likely and mutually exclusive, then the probability that A will occur denoted by p (A) is

p(A)= n/n+m = Number of outcomes favourable to occurrence of A / Total number of possible outcomes and the probability, that A will not occur, denoted by q (A) is

q(A) = m/n+m = Number of outcomes not favourable to occurrence of A / Total number of possible outcomes

For example, if we toss a coin, the probability of the head coming up is: p = 1/2, because the number of favourable event is 1, and the total number of possible outcomes is 2. The probability of head not coming up is:

$$q = 1/2$$

p and q of an event is equal to 1. (p+q = 1). Then, 1 - p = q, 1 - q = p.

Let us understand the Fundamental Concepts Used in Probability Theory

(i) Random Experiment

Probabilities are obtained for the outcomes of the situations, which are called random experiments. The term 'experiment' is used in Statistics in a much broader sence than in Chemistry or Physics. The tossing of a coin, for example, is considered a statistical experiment. An experiments has two properties: (a) each experiment has several possible outcomes and that can be specified in advance. (b)

We are uncertain about the outcome of each experiment. While tossing a coin, it can be specified that head or tail will turn up, but we are not certain whether the outcome of a particular toss will be head or tail. We use the word 'experiment' because the outcome is yet to be determined, whereas, the objective 'random' signifies that any particular outcome is uncertain.

(ii) Sample Space

A sample space of an experiment is the set (or collection) of all possible outcomes. The sample space for tossing a coin contains just two outcomes, head or tail; thus the sample space = (Head, Tail). The sample space of throwing a die will be (1, 2, 3, 4, 5, 6). Each possible outcome given in the sample space is called element or sample point. If two coins are to be tossed once, the four possible outcomes of this experiment will be:

Coin 2 / Coin 1	
H T	H T HH HT TH TT

The sample space of this experiment = (HH. HT, TH, TT).

If a pair of dice is to be cast once, the 36 possible outcomes of this experiment, will be:

Outcome of First Die	Outcome of Second Die 1	2	3	4	5	6
1	(1, 1)	(1,2)	(1, 3)	(1.4)	(1, 5)	(1, 6)
2	(2, 1)	(2,2)	(2, 3)	(2.4)	(2, 5)	(2, 6)
3	(3, 1)	(3,2)	(3, 3)	(3.4)	(3, 5)	(3, 6)
4	(4, 1)	(4,2)	(4, 3)	(4.4)	(4, 5)	(4, 6)
5	(5, 1)	(5,2)	(5, 3)	(5.4)	(5, 5)	(5, 6)
6	(6, 1)	(6,2)	(6, 3)	(6.4)	(6, 5)	(6, 6)

(iii) Event

An event is any element or point of the sample space in which we are interested. An event is called simple event, if it contains one element, if it is made up of more than sample point, it is called compound or complex event. A simple event is not decomposable, while a compound event can be decomposed into a number of disjoint simple events.

(iv) Mutually Exclusive Events

Mutually exclusive events are such events where the occurrence of one event prevents the possibility of the other to occur. In simple words, when several events are mutually exclusive, at the most one event may occur. A very simple example of a collection of mutually exclusive events is given by the coin toss. There are two possible events, a head or a tail. Since both events cannot occur on the same toss, they are mutually exclusive, the occurrence of one event rules out the occurrence of the other.

(v) Equally Likely Events

Events are said to be equally likely if after all relevant evidence has been taken into ac- count, one of them may not be expected rather than the other. For example, head and tail are equally likely events in tossing an unbiased or symmetrical coin.

(vi) Exhaustive Events

The events are said to be exhaustive if at least one of them necessarily occurs. In other words events are defined to be exhaustive if they between themselves exhaust all possible outcomes of the random experiment. For example, throwing of a die consists of 6 exhaustive events.

(vii) Independent Events

Events are said to be independent, if the occurrence of one does not affect the occurrence of any of the others. Two events are independent when they have no influence on each other. The result of the first toss of a coin does not affect the result of successive tosses at all.

(viii) Dependent Events

If the occurrence of the one event affects the happening of the other events, then they are said to be dependent events. For example, the probability of drawing a king from a pack of 52 cards is 4/54 or 1/12; if it happens that king is drawn and is not replaced in the pack, the probability of drawing again a king

would be 3/51. Thus, the outcome of the first event has affected the outcome of the second event. So they are dependent events.

(ix) Complementary Events

To any event A, there is an event denoted by 'not A' or A and called the complementary of A. A contains all the outcomes of the experiment which are not in A. Thus 'No head' or 'At least one head' are complementary events in two tossing of a coin.

From the concepts discussed above, it may be observed, that a probability will always be a number between 0 and 1 inclusively. This is because the numerator in the prob1a0b3ility fraction can never be negative nor can it be larger than the denominator. Two important observations follow from this definition. First, an event that is certain to occur will have the same value in both the numerator and the denominator, for the same events will result from all experiments. The probability in such a case will be 1. At the other extreme in impossible event's frequency ratio will always be 0 in the numerator, for such an event will occur in name of the experiments. Thus:

p (certain event) = 1, p (impossible event) = 0

Expressions of Probability

Probabilities can be expressed either as ratios, fraction or in percentages. For example, the probability of getting a head in a toss of coin can be expressed as 1/2 or .5 or 50%.

Illustration 3.1

(a) What is the chance of drawing a king in a draw from a pack of 52 cards?

Solution: Total number of cases that can happen = 52.

No. of favourable cases = total number of kings in a pack of cards = 4 The Probability (p) = 4/52 or 1/13.

(b) An urn contains two blue balls and three while balls. Find the probability of a blind man obtaining one blue ball in a single draw.

Solution: p = 2/(2+3) = 2/5.

(c) If two dice are thrown –

(i) What is the probability of throwing two sixes?

(ii) What is the probability of throwing a total of 9?

(iii) What is the probability of not throwing a total of 9?

Solution: The total number of outcomes in the throw of two dice will be 36.

Ist	2nd	Ist	2nd	Ist	2nd	Ist	2nd	Ist	2nd	Ist	2nd
1	1	2	1	3	1	4	1	5	1	6	1
1	2	2	2	3	2	4	2	5	2	6	2
1	3	2	3	3	3	4	3	5	3	6	3
1	4	2	4	3	4	4	4	5	4	6	4
1	5	2	5	3	5	4	5	5	5	6	5
1	6	2	6	3	6	4	6	5	6	6	6

(i) In the throw of two dice, two sixes can come only once, when the total number of outcomes will be 36. Hence the probability of coming of two sixes in the throw of two dice will be 2/36.

(ii) In the throw of two dice, total of 9 can come in this way : 3,6 ; 4 ,5; 5, 4; 6, 3. Hence, the probability of coming of a total of 9 in the throw of tw1o04dice will be 4/36 or .

(iii) The probability of not throwing a total of 9 in the throw of two dice will be 1 - 1/9 = 8/9.

(d) What is the probability that a vowel selected at random in any English book is an 'I'?

Solution: Total number of equally likely events = 5 Number of favourable events = 1 p = 1/5.

(e) What is the probability of a king in a pinochle deck?

(A pinochle deck consists of 2 aces. 2 kinds, 2 queens, 2 jacks, 2 tens and 2 nines of each suit. There are no cards of lower value).

Solution: Total number of cases = Total No. of cards = 48

Total number of favourable cases = Total number of king in the deck = 8

The p = 8/48 = 1/6.

(f) The ten digits 0 to 9 are stamped on 12 discs, there being one digit on a disc and the discs are thoroughly mixed in a box. If a disc is drawn at random, find the probabil- ity that the disc has an odd digit on it.

Solution: Total number of cases = 10.

Total number of favourable cases (1, 3, 5, 7 and 9) = 5

The p = 5/10 or 1/2 or 0.5.

(g) Find the probability of drawing a black card in a single random draw from a well- shuffled pack of ordinary playing cards.

Solution: Total number of outcomes = 52 No. of favourable outcomes = 26

Hence, p (drawing a black card) = 26/52 = 1/2.

(h) Find the probability of drawing a face card in a single random draw from a well- shuffled pack of ordinary playing cards.

Solution: There are 52 mutually exclusive equally likely outcomes.

The number of favourable outcomes (face cards - include the jack, the queen and the king in each) is 12.

Thus, p (drawing a face card) = 12/52 = 3/13.

By finding probabilities, for different events, a probability table can be constructed. In such a table, the probabilities of happening of all possible events can be seen simultaneously.

Illustration 3.2

There are 50 balls, each ball having two colours, one black or white and the other red, orange or green as shown in the following table:

	Red	Orange	Green	Total
Black	3	12	15	30
White	7	3	10	20
Total	10	15	25	50

(It means there are three balls, which are black and red, 12 balls are black and orange and so on.)

If of these balls, one ball is selected at random, find the probability of each type of ball being drawn up.

Solution:

Probability Table:

	Red	Orange	Green	Total
Black	0.06	.24	.30	0.60
White	0.14	.06	.20	0.40
Total	0.20	.30	.50	1.00

3.2.2 Empirical Approach to Probability

The classical or a priori approach to probability, while useful for solving problems involv- ing games of chance, suffers from serious difficulties, and does not provide answers to wide range of other types of problems. For example, it can tell the probability producing defective items in a production process. Such sort of questions can be answered with reference to empirical data. The probability of an event can be obtained on the basis of past records of the frequency distribution. For example, if a train comes daily. Past records show that in the last 365 days it was late on 13 days, then the probability of its late coming is (p) = 13/365.

According to Van Mises. "If an experiment be repeated a large number of times under essentially identical situations, the limiting value of the ratio of the number of times the event A happens to the total number of trials of the experiments as the numbers of trials increases indefinitely, is called the probability of the occurrence of A" Thus

$P(A) = m/n$

It is assumed that the limit is finite and unique.

Here,

m = no. of times an event A occurs

n = no. of times the experiment is performed

Illustration 3.3

(a) The following table gives a distribution of wages:

Weekly wages:	30-35	35-40	40-45	45-50	50-55	55-60	65-65	65-70
No. of workers:	9	108	488	230	112	30	16	7

An individual is taken at random from the above group. Find the probability

(i) his wages were under 40,

(ii) his wages were 55 or over, and

(iii) his wages were either between 45- 50 or 35-40.

Solution (i)

(i) Total wage earners = 9 + 108 +488 + 230 +122 + 30 + 16 + 7 = 1000

No of wage earners earning wages below 40 is = 9 + 108 = 117

Thus (p) = 117/1000 = .117

(ii) No. of wage earners earning wages over 55 is = 30 + 16 + 7 = 53 Hence, (p) 53/1000 = .053

(iii) No. of wage earners earning was between 45-50 or 35-40 is 230 + 108 = 338.

Hence (p) = 338/1000 = .338

(b) The manufactures of 'Bajaj' scooter give choice to their customers to have either a double seated scooter or a single seated scooter. On analysis of the booked orders for those scooters, they find that 75% of their customers are men and 25% women. 80% of the men customers prefer double seated scooters and rest one seated. 90% of their women customers prefer one seated scooters and rest two seated scooters. In what proportion, the manufacturers should manufacture these two scooters?

Solution:

Men customers preferring tow seater = .75 x .8 = .600

Women customers preferring two seater = .25 x .1 = .025

Total = .625

Men customers preferring one-seater = .75 x .2 = .150

Women customers preferring one-seater = .25 x .9 = .225

Total = .375

Ratio: .625 : .375 = 5:3

(c) In a sample of 100 radios it was found that:

No. of defects	0	1	2
No. of Radios	10	85	5

What is the probability that a radio selected at random will have zero defects?

Solution (iii)

10/100 = 1/10 or 0.1 0r 10%

3.2.3 Subjective Approach to Probability

The subjective or personality approach to probability is of recent origin. According to this concept, the probability of an event is the degree of belief or degree of confidence placed in the occurrence of an event by a particular individual based upon the evidence available to him. This evidence may consist of relative frequency of occurrence of data or any other quantitative or qualitative information. To forecast the demand, predicting price etc. is done on the basis of subjective probabilities. The probability is determined be- tween 0(0 = impossible) and 1(1 = certain event).

Illustration 3.4

(a) A job applicant assigns probabilities as follows:

The probability, p(A), of being offered a job at a company. A is 0.6; the probability, p(B); of being offered a job at a company B is 0.5; the probability of being offered a job at both companies is 0.4. What, consequently, is the probability of being offered a job with at least one of the two companies?

Solution:

	B	Given B’	Total			Total B	Completed B’	Total
A	0.4	—	0.6		A	0.4	0.2	0.6
A’	—	—	—		A’	0.1	0.3	0.4
Total	0.5	—	1.0	Total		0.5	0.5	1.0

The probability of his getting job at least in one of the companies, i.e.; p(AB+A'B+AB') = 0.4+0.1+0.2 = 0.7.

(b) You have noticed that your officer is happy on 60 percent of your calls, so you assign a probability of his being happy on your visit as 0.6 or 6/10. You have no- ticed also that if he is happy, he accedes to your requests with a probability of 0.4 whereas if he is not happy, he accedes to the requests with a probability of 0.1. You call one day and he accedes to your request. What is the probability of his being happy?

Solution:

	Probability Table		
	Happy	Not happy	Total
Request Accepted	.24	.04	.28
Request not Accepted	.36	.36	.72
Total	.6	.4	1

(p) of being happy and accepting the request = .6 x .4 = 0.24

(p) of not being happy and accepting the request = .4 x .1 = .04

The chances of his accepting the request = 0.28

and the chances of his accepting the request when he is happy = .24

Hence, the probability of his being happy having accepted the request = .24/.28 or 6/7 or 0.857.

3.2.4 Modern Approach to Probability

In this approach no precise definition of probability is given, but the theory is based on certain axioms or postulates. The axioms are:

To every event A, there corresponds a real value P(A) called probability of the happening of the event A, which satisfies the following three axioms:

(i) $0 \leqslant P(A) \leqslant 1$:

(ii) $P(S) = 1$, $P(\phi) = 0$ S = certain event

ϕ = impossible event

(iii) If $A_1, A_2 \ldots\ldots$ An are mutually exclusive events, then

$(A_1$, or A_2 or……or $A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$ in particular,

P(A) = The sum of the probabilities of simple event, comprising the event A = Number of sample points in A / Total no. of sample points in S

3.3 Probability Theorems for Problems Solving

The solution to many problems involving probabilities requires a thorough understanding of some of basic rules that govern the manipulation of probabilities. They are generally called probability theorems. Let us discuss them in detail:

(1) Addition Theorem: The theorem is defined as follows:

"If two events are mutually exclusive and the probability of the one is p1 while that of the other is p2, the probability of either the one event or the other occurring is the sum p1+p2"

Proof of the Theorem: If an event A can happen in 'a 1' ways and B in 'a 2' ways, then the number of ways in which either event can happen is 'a 1 +a 2'. If the total number of possibilities is 'n', then by definition the probability of either the first or the second event happening is

a 1 +a 2/n = a 1/n +a 2/n

Since a 1/n = P(A) and a 2/n = P(B) Hence P(AorB) = P(A) + P(B).

The theorem can be extended to three or more mutually exclusive events. Thus,

P(A or B or C) = P(A) + P(B) + P(C)

For example, the probability of getting spot (1) in a throw of a single die is 1/6,

The probability of getting spot (3) is also 1/6 and the probability of getting spot (5) too is 1/6. The probability of getting and odd number (1, 3 and 5) in a throw of a single die will be the addition of their respective probabilities, that is, 1/6+1/6+1/ 6 = 3/6 or 1/2.

The addition theorem will hold good only if:

(i) Items are mutually exclusive,

(ii) Mutually exclusive items belong to same set.

Illustration 3.5

(a) A bag contains 4 white, 2 black, 3 yellow and 3 red balls. What is the probability of getting a white or red ball at random in a single draw of one?

(b) A card is drawn at random from an ordinary pack of 52 playing cards. Find the probability that a card drawn is either a spade or the ace of diamonds.

Solution (a):

The probability of getting one white ball = 4/12

The probability of getting one red ball = 4/12

The probability of one white or red ball = 4/12+3/12 = 7/12 or 7/12x100 = 58.3%.

Solution (b):

The probability of drawing a spade = 13/52.

The probability of drawing and ace of diamonds = 1/52.

Probability of drawing a spade or an ace of diamond = 13/52 + 1/52 = 14/52 A total of 7 can come in 6 different ways (1/6, 2/5, 3/4, 4/3, 5/2, 6/1)

A total of 11 can come in 2 different ways (5/6, 6/5) The probability of getting a total of 7 = 6/36 or 1/6 The probability of getting a total of 11 = 2/36 or 1/18

The probability of getting either 7 or 11= 1/6 + 1/18 = 4/18 or 2/9.

The addition theorem will hold good only if the events are mutually exclusive. If events contain no sample point in common, then some adjustment is necessary under such a case:

p [(A) or (B)] = p (A)+p(B)-p(A and B)

The following example will make it clear.

A bag contains 25 balls, numbered from 1 to 25, one is to be drawn at random. Find the probability that the number of the drawn ball will be a multiple of 5 or 7.

The probability of the number being multiple of 5 (5, 10, 15, 20, 25) = 5/25. The probability of the number being multiple of 7 (7, 14, 21) = 3/25

Thus the probability of the number being a multiple of 5 or 7 will be = 5/25+3/25= 8/25.

In the above illustration, find the probability that the number is a multiple of 3 or5: The probability of the number being multiple of 3 (3,6 9, 12, 15, 18,21, 24) = 8/25

The probability of the number being multiple of 5 (5, 10, 15, 20, 25) = 5/25.

Joint probability 8/25+5/25 = 13/25; but this answer is wrong, because item No. 15 is not mutually exclusive. Hence the correct probability will be

= 8/25+5/25-1/25 = 12/25.

Hence, p (A+B) + p (A) + p(B)-p(AB). The following diagram will make it clear.

Similarly, when three events are not mutually exclusive, then:

p(A+B+C) =p(A)+p (B)+p (C) -p (AB) -p (AC)-p (BC)+p (ABC)

Illustration 3.6

What is the probability of drawing a black card or a king from a pack of ordinary playing cards?

Number of black cards		number of kings		number of black kings
26	+	4	-	2

Hence, (p) = 26/52+4/52-2/52 = 28/52.

It is also essential that mutually exclusive items must belong to the same set. To illustrate this point, let us look at the following example:

Suppose the probability of a man dying between his 40th and 41st birth days is 0011, and the probability of his marrying between his 41st and 42nd birthdays is 0.009. These events are mutually exclusive but it cannot be said that the probability of a man dying in his 40th year and of marrying in his 41st year is .011+.009 = .02. These two events do not belong to the same set.

(2) Multiplication Theorem: According to this theorem. “If two events are mutu- ally independent, and the probability of the one is P1 while that of the other is P2 the probability of the two events

occurring simultaneously is the product of P1 and P2". For example, the probability of head coming up in a toss of a coin is 1/2 and the probability of 4 coming in a throw of a die is 1/6. If a coin and a die are thrown together, the probability of head coming up in the toss of coin and 4 coming up in the throw of a die will be 1/2x1/ 6 =1/12.

Proof of the Theorem: If an event A can happen in 'n_1' ways of which 'a_1' are successful and the event B can happen in 'n_2' ways of which 'a_2' ways are successful, we combine the successful events of both A and B events where the total number of successful happening is '$a_1 \times a_2$'. Similarly, the total number of possible cases is '$n_1 \times n_2$'. Then by definition, the probability of occurrence of both event is,

$$a_1 \times a_2/n_1 \times n_2 = a_1/n_1 \times a_2/n_2$$

Since $a_1/n_1 = P(A)$ and $a_2/n_2 = P(B)$

Hence, P(A and B) = P(A) x P(B).

The theorem can be extended to three or more independent events. Thus,

P(A and B and C) = P(A) x P(B) x P(C)

Illustration 3.7:

(a) What is the probability of throwing two 'fours' in two throws of a die?

(b) What is the probability of getting all the heads in four throws of a coin?

(c) Suppose it is 9 to 7 against a person A who is now 35 years of age lying till he is 65 and 3 to 2 against a person B, now 45 living till he is 75; fund the chance that one at least of these persons will be alive 30 yearshence.

(d) A problem in statistics is given to three students A, B, C, whose chances of solving it are 1/2, 1/3, 1/4 respectively. What is the probability that the problem will besolved?

(e) A Helicopter is equipped with three engines that operate independently. The probability of an engine failure is 0.01. What is the probability of successful fight if only one engine is needed for the successful operation of the aircraft?

Solution:

(a) The probability of a 'four' in first throw = 1/6

The probability of a 'four' in second throw = 1/6 The probability of two 'fours' = 1/6 x 1/6 = 1/36.

(b) The chance of getting head in the Ist throw = 1/2

The chance of getting head in the 2nd throw = 1/2

The chance of getting head in the 3rd throw = 1/2

The chance of getting head in the 4th throw = 1/2

Thus the probability of getting heads in all the throws;

=1/2x1/2x1/2x1/2 = 1/16.

(c) The chance that A will die within 30 years is 9/16 and the chance that B will die within 30 years is 3/5. The events are independent, therefore, the chance that both will die is 9/16 x 3/5 = 27/80.

Then chance that both will not be dead i.e., at least one will be alive is 1-27/80

= 53/80.

(d) Probability that student A will fail to solve the problem = 1 - 1/2 = 1/2

Probability that student B will fail to solve the problem = 1 - 1/3 = 2/3

Probability that student C will fail to solve the problem = 1 - 1/4 = 3/4 Since the events are independent, the probability that all the students A, B, C will fail to solve the problem = 1/2 x 2/3 x 3/4 = 1/4.

So, the probability that the problem will be solved = 1 - 1/4 = 3/4. This problem can also be solved in the following way:

	Condition	Probability
(i)	A solves, B solves C solves	= 1/2 x 1/3 x 1/4 = 1/24
(ii)	A solves, B solves, C fails to solve	= 1/2 x 1/3 x 3/4 = 3/24
(iii)	A solves, B fails to solve, C solves	= 1/2 x 2/3 x 1/4 = 2/24
(iv)	A fails to solve, B solves, C solves	= 1/2 x 1/3 x 1/4 = 1/24

(v) A solves, B fails to solve, C fails to solves = 1/2 x 2/3 x 3/4 = 6/24

(vi) A fails to solves, B solves, C fails to solve = 1/2 x 1/3 x 3/4 = 3/24

(vii) A fails to solves, B fails to solve, C solves = 1/2 x 2/3 x 1/4 = 2/24

(viii) A fails to solves, B fails to solve, C fails to solve = 1/2 x 2/3 x 3/4 = 6/24.

The problem is solved in all the conditions, except that of (viii). If the probabilities of (i) to (vii) are added, that will give the probabilities of problem being solved. The total comes to 18/24 or 3/4.

(e) Since the flight is unsuccessful only when all the three engines fail, then the prob ability of unsuccessful flight is: .01 x .01 x.01 = .000001.

The probability of successful flight = 1 - .000001 = .999999.

Illustration 3.8

Five cards are to be drawn in succession and without replacement from an ordinary deck of playing cards.

(a) What is the probability that there will be no ace among the five cards drawn?

(b) What is the probability that the first three cards are aces and the last two cards are kings?

(c) What is the probability that only the first three cards are aces?

(d) What is the probability that an ace will appear only on the fifth draw?

Solution:

(a) The probability that there will be no ace among the five cards:

p = 48/52 x 47/51 x 46/50 x45/49 x 44/48 = 205476480/311875200.

(b) The probability that the first three cards are aces and the last two cards are kings:

p = 4/52 x 3/51 x 2/50 x4/49 x 3/48 = 288/311875200.

(c) The probability that only the first three cards are aces:

p = 4/52 x 3/51 x 2/50 x48/49 x 47/48 = 54144/311875200.

(d) The probability that an ace will appear only on the fifth draw:

p = 48/52 x 47/51 x 46/50 x45/49 x 4/48 = 18679680/31875200.

The multiplication theorem will hold good only if the events belong to the same set. In order to show the importance of this fact, Moroney in his book "facts from Figures" gives an interesting example. He observes, "Consider the case of a man who demands the simultaneous occurrence of many virtues of an unrelated nature in his young lady. Let us suppose that he insists on a Grecian nose, platinum-blonde hair, eyes of odd113 colours - one blue, one brown, and finally a first class knowledge of statistics. What is the probability that the first lady he meets in the street will put ideas of marriage into his head? It is difficult to apply multiplication theorem in this case, because events do not belong to the same set.

(3) Conditional Theorem: If sub-event are not independent, and the nature of dependence is known, we have the theorem of conditional probabilities. This theorem is more or less corollary of the multiplication theorem. The theorem is that the probability that both of two dependent sub-events can occur is the product of the probability of the first sub-event and the probability of the second after the first sub-event has occurred. In notation p(A and B) = p (A) x p(B/A) is the conditional probability of B when A has already happened. For example, if out of a pack of cards shuffled or each time 'king' turns out first and the card is not restored, then in a second reshuffling the prob ability of 'king' turning up again - 4/52 x 4/51 = 12/2652, since there are 4 kings at the first shuffle of 52 cards and 3 kings only at the second shuffle of 51 cards. The term 'condition probabilities' is often known as probabilities due to partial exhaustion of a sample space.

(4) Bayes' Theorem: Probabilities can be revised when new information pertaining to a random experiment is obtained. The notion of revising probabilities is a familiar one, for all of us, even to those with no previous experience in calculating probabilities - have lived in an environment ruled by whims of chance and have made informal probability judgements. We do also intuitively revise these probabilities upon observing certain facts and change our actions accordingly. Our concern for revising probabilities arises from a need to make better use of experimental information. This is referred to as Bayes' Theorem after the Reverend Thomas Bayes, who proposed in the eighteenth century, that probabilities be revised in accordance with empirical findings.

Quite often the businessman has the extra information on a particular event or proposition, either through a personal belief or from the past history of the event. Probabilities assigned on the basis of personal experience, before observing the outcomes of the experiment are called prior prob- abilities.

For example, probabilities assigned to past sales records, to past number of defectives produced by a machine, are examples of prior probabilities. When the probabilities are revised with the use of Bayes; rule, they are called posterior probabilities. Bayes' theorem is very useful in solving practical business problems in the light of additional information.

Suppose, a random experiment having several mutually exclusive events E1, E2…… and the probabilities of each event P(E1), P(E2) have been obtained. These probabilities are referred to as prior probabilities, because they represent the chances that events before the results from empirical investigation are obtained. The investigation itself may have several possible outcomes, each statistically dependent upon Es. For any particularly result which we may designate by the letter R, the conditional probabilities P(R/E1), P(R/E2) are often available.

The result itself serves to revise the event probabilities upward or downward. The resulting values are called posterior probabilities since they apply after the information result has been learned. The posterior probability values are actually conditional probabilities of the form P (E1/R), P (E2/

R) that may be found according to Bayes' Theorem. The posterior probability of E, for a particular result R of an empirical investigation may be found from:

$P(E2/R) = P(E2)\ P(R/E1)/ [P(E1)\ P(R/E1) + P(E2)\ P(R/E2)]$

Illustration 3.9

Box 1 contains three defective and seven non-defective items, and Box 2 contains one defective and nine non-defective items. We select a box at random and then draw one item from the box.

(a) What is the probability of drawing a non-defective item?

(b) What is the probability of drawing a defective item?

(c) What is the probability that box 2 was chosen, given a defective item is drawn?

Solution:

P(B1) = Probability that box 1 is chosen = 1/2,

P(B2) = Probability that box 2 is chosen = 1/2,

P(D) = Probability that a defective item is drawn,

P(ND) = Probability that a non-defective item is drawn.

(a) P(ND) = P (Box 1 and non-defective) + P(Box 2 and non-defective)

= (1/2 x 7/10) + (1/2 x 9/10) = 16/20

(b) P(D) = P (Box 1 and defective) + P(Box 2 and defective) or

= P(D) = (1/2 x 3/10) + (1/2 x p/10) = 4/20

(c) By Bayes' theorem

P(B1/D) = PB1 and D) /P(D) = 3/20 / 4/20 = 3/4.

P(B1) and P(B2) are called prior probabilities and P(B1/D) and PB2/D) are called posterior probabilities. The above information is summarized in the following table:

Event	Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
B1	½	3/10	3/20	3/4
B2	½	1/10	1/20	1/4
Total			**4/20**	**1.0**

Illustration 3.10

A box contains four fair dices and one crooked die with a loaded weight which makes the six-face appear on two-thirds tosses. You are asked to select one, die at random and toss it. If the crooked die is indistinguishable from the fair die and the result of your toss is a six-face; what is the probability that you tossed the crooked die?

Solution:

P (Fair dice) = 4/5

P (Crooked die) = 1/5

The probability of tossing a six face in a fair die = 1/6

The probability of tossing a six face in a crooked die = 2/3

The probability of tossing a six face when die is crooked = 1/5 x 2/3 = 2/15.

The probability of tossing a six face when die is fair = 1/5 x 1/6 = 4/30.

The probability of tossing a six face = 2/15 + 4/30 = 4/15

The posterior probability that die tossed is crooked is: 2/15 / 4/14 = 1/2 or 50%.

Illustration 3.11

Urn A1 contains 8 black and 2 white marbles. Urn A3 contains 3 black and 7 white marbles, and urn A3 contains 5 white and 5 black marbles. A fair die is to be cast. If the die turns up 1, 2 or 3 then a marble will be selected from A1. If the die turns up 4 or 5 a marble will be selected from A2. Finally, a marble will be selected from A3. If the die turns up 6. Given that the marble selected is black, what is the probability that the marble was from urn A2?

Solution:

Probability of marble being chosen from urn A1 = 3/6 Probability of selecting a black marble from A1 = 8/10

Hence, the joint probability of 1, 2 or 3 coming up in the fair die and then drawing a black marble from A1 = 3/6 x 8/10 = 24/60.

Similarly, the probability of 4 or 5 turning up in the die and drawing a black marble from urn A2 = 2/6 x 3/10 = 6/60 and

the probability of 6 turning up in the die and drawing a black marble from urn A3 = 1/6 x 5/10 = 5/60

The probability of drawing a black marble from any of these urn is 24/60 + 6/60 + 5/60 = 35/60.

Assuming that the marble selected is black, the probability that the marble was chosen from urn A2 is: P 6/60 / 35/60 = 6/35.

Illustration 3.12

Urn A contains 6 green and 4 red marbles, and urn B contains 2 green and 7 red marbles. A marble is to be selected at random from A and placed in B. One marble is then selected from B. Given that the marble selected from B is green, what is the probability that the marble selected from A will also be green?

Solution

Probability of marble selected from B is green, if the marble selected from A and p1aced in B is green = 6/10 x (2+1)/(9+1) = 6/10 x 3/10 = 18/100. Probability of marble selected from B is green; if the marble selected from A and placed in B is red

= 4/10 x 2/9+1 = 4/10 x 2/10 = 8/100.

The joint probability of green marble selected from B = 18/100 + 8/100 = 26/100.

The probability, given that the marble selected from B is green, the marble selected from A will also be green.

=18/100 / 26/100 = 18/26.

Illustration 3.13

In a factory, machines M1, M2 and M3 manufacture respectively, 30, 30 and 40 percent of the total output. Of their output 1, 3, and 2 percent are defective items. An item is drawn from day's output and is found defective. What is the probability that it was manufactured by M1 by M2, by M3?

Solution:

(P) that an item is manufactured by M1 = 30/100 = .3, and the item is defective:

.3 x .01 = .003.

(P) that an item is manufactured by M2 = 30/100 = .3, and the item is defective:

.3 x .03 = .009.

(P) that an item is manufactured by M3 = 40/100 = .4, and the item is defective:

.4 x .02 = .008.

Probability of defective item = .003+.009+.008 = .02

Probability that the defective item is manufactured by M1 = .003/.02 = or 3/20. Probability that the defective item is manufactured by M2 = .009/.02 = or 9/20. Probability that the defective item is manufactured by M3 = .008/.02 = or 8/20.

Illustration 3.14

A can hit a target 3 times in 5 shots, B 2 item in 5 shots, C 3 times in 4 shots. They fire a volley. What is the probability that 2 shots hit?

Solution:

Fire a volley means that A, B and C all try to hit the target simultaneously. Two shots hit the target in one of the following ways:

(a) A and B hit and C fails to hit.

(b) A and C hit and B fails to hit.

(c) B and C hit and A fails to hit.

The chance of hitting by A = 3/5 and of not hitting by him = 1- 3/5=2/5

The chance of hitting by B = 2/5 and of not hitting by him = 1- 2/5=3/5

The chance of hitting by C = 3/4 and of not hitting by him = 1- 3/4=1/4

The probability of (a) = 3/5 x 2/5 x 1/4 = 6/100

The probability of (b) = 3/5 x 3/4 x 3/5 = 27/100

The probability of (c) = 2/5 x 3/4 x 2/5 = 6/100

Since (a), (b) and (c) are mutually exclusive events, the probability that two shots hit

6/100 + 27/100 + 12/100 = 45/100 = 9/20 or 9/20 x 100 = 45%

The classical or a prior probability measures have two very interesting characteristics. First, the objects referred to as fair coins true dice or fair deck of cards are abstracting in the sense that no real world object exactly possesses the features postulated.

Secondly, in order to determine the probabilities, no coins had to be tossed, no dice rolled nor cards shuffled. That is no experimental data were required to be collected; the probability calculations were based entirely on logical prior (thus a priori) reasoning.

It may be possible that the results of a few trials of an experiment may be different than the expected on the basis of probability. If a coin is tossed 10 times, it may be that head may turn up 7 times and tail 3

times whereas, according to the prior probability the head should turn 5 times and tail also 5 times. But in 500 or 1000 trials, the results may be much nearer to the probable results.

3.4 Permutation and Combinations in the Theory of Probability

Knowledge of permutaions and combinations is essential to solve the problems related to probability determination. So, we have discussed these concepts hereunder:

Permutations: Sometimes we are interested in the total number of different ways in which items can be arranged so that the order of components is important, yet no two arrangements are similar. Arrangements of this sort are called permutations. For ex- ample, if seven alphabets - A, B, C, D, E, F, G, are to be arranged by taking two letters at a time, but under no circumstances may an arrangement contain the same 2 letters (like AA, or BB etc.) then the following permutations are possible:

AB	AC	AD	AE	AF	AG
BA	BC	BD	BE	BF	BG
CA	CB	CD	CE	CF	CG
DA	DB	DC	DE	DF	DG
EA	EB	EC	ED	EF	EG
FA	FB	FC	FD	FE	FG
GA	GB	GC	GD	GE	GF

Hence, there are 7 x6 = 42 permutations,

Thus, following formula can give the number of permutations

Perm. = n(n - 1)

If 26 letters are to be arranged in this manner, the total number of ways will be 26 (26 - 1) = 650.

The permutation can be shown in a tree-diagram also. For example, three chairs x, y and z can be arranged n(n - 1)(n - 1) or 3(3 - 1)(3 - 2) = 6 ways. The tree diagram shows this:

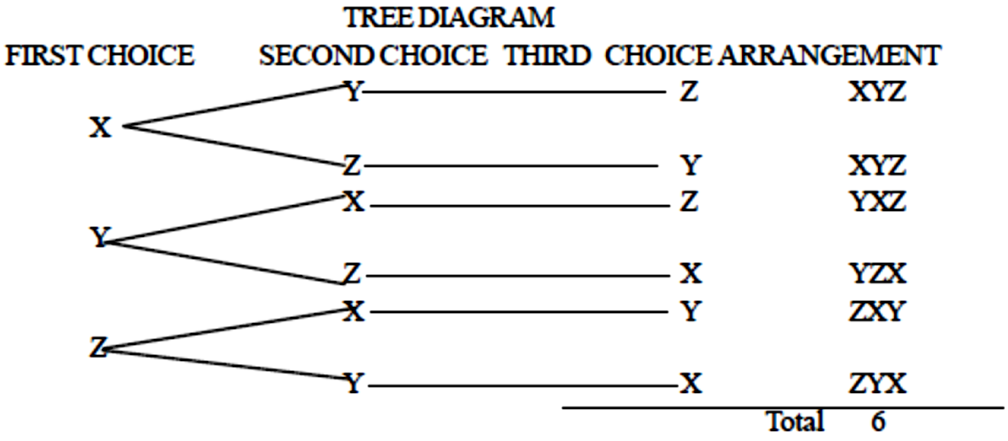


Illustration 3.15:

If a man has the choice of traveling between Hisar and Delhi by 8 trains, in how many possible ways he can complete the return journey, using a different train in each direction?

Solution:

For the outward journey he has the choice of using all the 8 trains. Having completed the outward journey, he will be left with only 7 trains to complete the return journey.

Thus, the total number of ways in which he can complete the journey are 8(8 - 1) = 56.

Some general rules regarding permutation are as follows:

For finding permutation of doing 'n' function in 'r' ways, the formula is Prem. = n(n -1) (n - 2) x (n - 3) x (n - 4)….

This formula can also be written as follows: P = n !*/(n - r) !

Illustration 3.16

(a) If a person is given one cup of coffee of each of 5 brands and asked to rank these according to preference. How many possible ranking can there be?

n Pr = n ! / (n - r) ! or 5 !/(5 -5) = 5 x 4 x 3 x 2 x1/1 = 120 It can also be calculated:

Prem. = n(n - 1) (n - 2) (n - 3) (n - 4) = 5 x 4 x 3 x 2 x 1=120

(b) There are six doors in a room. Four persons have to enter it. In how many ways they can enter from different doors?

n(n - 1) (n - 2) (n - 3) n = 6, r = 6

=(6 - 1) (6 - 2) (6 - 3) 4p4 = n !/ (n - r) !

= 6 x 5 x4 x 3 = 360 = 6 ! / (6 - 4) ! 6 x 5 x 4 x 3 x 2 x 1 / 2x 1 = 6 x 5 x 4 x 3 =360

(c) In how many ways first, second and third prizes can be distributed to three of 10 competitors?

n = 10, r = 3

p3 = 10 ! / (10 - 3) = 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1 / 7 x 6 x 5 x 4 x 3 x 2 x 1 =

10 x 9 x 8 = 720 ways.

(d) Four strangers board a train in which there are 6 empty seats. In how many different ways can they be seated?

n = 6, r = 4

6p4 = n !/ (n - r) !/ 6 !/ (6 - 4) ! = 6 x 5 x 4 x 3 x 2 x 1 / 2 x 1

= 6 x 5 x 4 x 3 = 360 ways

(e) In how many ways can 12 seats be occupied by 6 women? n = 12, r = 6

13p4 = n !/ (n - r) ! = 12 !/ (12 - 6) ! = 12 x 11 x 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2

x 1 / 6 x 5 x 4 x 3 x 2 x 1 = 12 x 11 x 10 x 9 x 8 x 7 = 665280.

(f) How many of officers can we possibly have if we have a set of 10 persons and have to different sets fill the offices of chairman, Dy. Chairman, Secretary and Treasurer.

p4 = 10 ! / (10 - 4) = 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1 / 6 x 5 x 4 x 3 x 2 x 1= 5040.

If there are n things in which p are of one kind, q are of the other kind and r of a third kind, then the number of permutations will be:

Prem. = n ! / P x q x r

Illustration 3.17

(a) In how may ways 12 students of MBA (DE) be allotted to three tutorial groups of 2, 4 and 6 respectively?

n = 12, p = 2, q = 4, r = 6

n ! / p !q!r! = 12 ! / 2!4! 6!

= 12 x 11 x 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1/ (2 x 1) (4 x3 x 2 x 1) (6 x 5 x 4 x 3 x 2 x 1)

= 12 x 11 x 10 x 9 x 8 x 7 / 4 x 3 x 2 x 1 x 2 x 1 = 13860.

(b) In how may ways can the letters of the word 'BASKET' be arranged? There are 6 letters.

Hence Perm. = 6 ! = 6 x 5 x 4 x 3 x 2 x 1 = 720 ways.

(c) How many ways can the letter of the words 'BETTER' be arranged?

In this letter E comes twice. 'T' comes twice, 'B' and 'R' come only once.

Hence, Perm. = n ! / p !q!r! = 6 ! / 2!2!1 = 6 x 5 x 4 x 3 x 2 x 1/ (2 x 1) x (2 x 1) x 1 = 720/

4=108.

(c) In how many ways can the letters of the words 'MACMILLAN', 'BANANA'AND 'STATISTICALLY' can be arranged?

'MACMILLAN' N= 9

A comes 2 times M comes 2 times L comes 2 times others come only once

Permutation = 9 ! / 2! 2! 2! = 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1/ (2 x 1) (2x 1) (2 x 1) = 15120.
'BANANA' n = 6 A = P = 3

N = q = 2

Permutation = n! / p! q! = 6 ! / 3! 2! = 6 x 5 x 4 x 3 x 2 x 1/ (3 x 2 x 1) (2x 1) = 60.
'STATISTICALLY' n = 13 T = p = 3

S = q = 2

A = r = 2 I = l = 2 L = k = 2 others come only one:

Permutation = n ! / p!q!r!l!k! = 13 ! / 3! 2! 2! 2! 2!

= 13 x 12 x 11 x 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3

x 2 x 1/ (3 x 2 x 1) (2 x 1) (2 x 1) (2 x 1) (2 x 1) = 64864800.

3.5 CHECK YOUR PROGRESS

1. The sum of all probabilities equal to _____.

2. The probability of drawing an ace card from a deck of cards is _____.

3. A dice is thrown. Find the probability of getting an even number.

4. If the probability of winning a game is 0.3, then probability of losing it is _____.

5. The sum of the probabilities of all elementary events of an experiment is p, then p equals to _____.

3.6 SUMMARY

The word probability or chance is very common in day to day life of human being. Probability is especially important in statistics because of the many principles and procedures that are based on this concept. Indeed, probability plays a special role in all our lives, and has an everyday meaning. Sometimes we hear phrases like: 'You had better take an umbrella because it is likely to rain.' 'His chances of winning are pretty small.' It is very likely that it may rain by the evening. You are probably right.' or 'There are fifty-fifty chances of his passing the examination.' In each of these phrases an idea of uncertainty is acknowledged. Goethe remarked that, "There is nothing more frightful than action in ignorance." Reasoning in terms of probabilities is one weapon by which we attempt to reduce this uncertainty or ignorance. The use of word 'probability' in statistics, however is somewhat different. It is more precise than what it means in popular usage. In statistics, a probability is a numerical value that measures the uncertainty that a particular event will occur. There are basically three methods of measuring probabilities. They represent different conceptual approaches. They are: Classical Approach, Empirical Approach, Subjective Probability, and Modern Approach.

3.7 KEYWORDS

Random experiment: A process of obtaining information through observation or measurement of a phenomenon whose outcome is subject to chance.

A simple event: The basic possible outcome of an experiment, it cannot be broken down into simpler outcomes.

Sample space: The set of all possible outcomes or simple events of an experiment.

Event: Any set of outcomes of an experiment, a subset of the sample space.

Probability: A numerical measure of the likelihood of occurrence of an uncertain event. Collectively exhaustive events: The list of events that represents all possible experimental outcomes.

Mutually exclusive events: Events which cannot occur together, that is, events having no sample points in common (disjoint).

Conditional probability: The probability of an event occurring, given that another event has occurred.

Joint probability: The probability of two events occurring together or in succession.

Marginal probability: The unconditional probability of an event occurring.

Statistical dependence: The condition when the probability of occurrence of an event is dependent upon, or affected by, the occurrence of some other event.

Statistical independence: The condition when the probability of occurrence of an event has no influence on the probability of occurrence of any other event.

Posterior probability: A revised probability of an event obtained after getting additional information.

Baye's theorem: A method to compute posterior probabilities (conditional probabilities under statistical dependence).

3.8 SELF ASSESSMENT QUESTIONS

1. Define probability and explain the importance of this concept in statistics.

2. Explain what do you understand by term ‘probability’. State and prove the addition and multiplication theorems of probability.

3. Explain the concept of independence and mutually exclusive events in probability. State theorems of total and compound probability.

4. Define probability and enunciate the Multiplication Law of probability, giving suitable examples.

5. What are the different schools of thought on the interpretation of probability? How does each school define probability? Explain with suitable examples.

6. (a) When are the events said to be independent in the probability sense? Give examples of dependent and independent events.

   (b) Differentiate between the circumstances when the probabilities of two events are : (i) added, and (ii) multiplied.

7. (a) If A and B are events, define the compound events: A+B, (i.e., the union of the two events and A x B, i.e., the intersection of the two events. Prove that

   $$p(A+B)= p(A)=p(B) - p(AB)$$

   and establish a similar rule for p(A+B+C). State the general result for n such events.

   (b) When a soldier fires at target, the probability that he hits the target is 1/2 for soldier A, 1/2 for soldier B, 2/3 for soldier C, and 1/12 for soldier D. If all the four soldiers A, B, C and D fire at the target simultaneously, calculate probability that the target is hit by some one or more.

   [(b) 373/648.

8. Explain why there must be mistake in the following statement:

" A quality control engineer claims that the probability that a large consignment of glass bricks contains 0, 1, 2, 3, 4 or 5 defectives are .11, .23, .37, .16, .09 and .05 respectively. [The total of probabilities of all mutually exclusive events cannot exceed 1. Here it is 1.01].

9. One bag contains 4 white balls and 2 black balls. Another contains 3 white balls and 5 black ball. If one ball is drawn from each bag, find the probability that (a) both are white,

(b) both are black, and (c) one is white and one is balck. [(a) 1/4, (b) 5/24, (c) 13/24.

10. A person is known to hit the target in 3 out of 4 shots, whereas another person is known to hit 2 out of 3 shots. Find the probability of the target being hit at all when they both try. [11/12]

11. The odds are 7 to 5 against A, a person who is now 30 years old living till he is 70 years and the odds are 2 to 3 in favour of B who is now 40 years of age living till he is 80 years. Find the chance that one at least of these two persons will be alive 40 years hence. [13/20]

12. It is given that in two towns the number of rainy days in a year are respectively 20 and 30. What is the probability that on a particular day in that year there is (a) no rain in both towns,

(b) rain in one town only, (c) rain in both towns.

[(a) 69/73 x 67/73, (b) 4/73 + 6/73, (c) 4/73 x 6/73]

3.9 ANSWERS TO CHECK YOUR PROGRESS

1. One

2. Probability = 1/13

3. Probability = ½

4. Probability = 0.7

5. p = 1

3.10 REFERENCES/SUGGESTED READINGS

Hooda, R P: Statistics for Business and Economics, 3rd Edition, MacMilan India Ltd.,

New Delhi.

Gupta, S P: Statistical Methods, 7th Edition, Sultan Chand and Sons, New Delhi. Bhardwaj, R S: Business Statistics, Excel Book, New Delhi.

Murray R. Spiegel and Larry J. Stephens, Statistics, 3rd Edition, TMH, New Delhi.

Viswanathan, PK, Business Statistics, First edition, Pearson Education (Singapore) Ltd., Delhi.

Lesson: 4	**Author:** Dr. Karam Pal Singh
Updated By: Ms. Chand Kiran	**Vetter:** Prof. B. S. Bodla

PROBABILITY DISTRIBUTION

STRUCTURE

4.0 LEARNING OBJECTIVES

After reading this lesson, you must be able to-

- Understand meaning and different types of probability distributions.
- Calculate probability involving in the Binomial, or Poisson or Normal distributions.
- Fitting the probability distributions under various cases.

4.1 INTRODUCTION

The probability distribution of a random variable may be: Probability listing of outcomes and probabilities which can be obtained from a mathematical model representing some phenomenon of interest; an empirical listing of outcomes and their observed relative frequencies; and a subjective listing of outcomes associated, with their subjective or 'contrived' probabilities representing the degree of conviction of the decision-maker as to the likelihood of the possible outcomes. In this lesson our main concern would be with the first kind of probability distribution.

4.1.2 MEANING AND UTILITY OF PROBABILITY DISTRIBUTIONS

The observed frequency distributions are based on observation and experimentation. For example, we may study the marks of the students of a class and classify the data in the form of a frequency distribution as follows:

Marks	No. of Students
0-20	20
20-40	24
40-60	50
60-80	30
80-100	16
Total	140

The above example clearly shows that the observed frequency distributions are obtained by grouping data. They help in understanding properly the nature of data. For instance, a shoe-maker must know something about distributions of sizes of his potential customers' feet, otherwise he may be stuck with shoe sizes that he cannot sell. Similarly, an owner of a restaurant must know about the people's likes and dislikes of various foods, otherwise he might find problems in running the business.

As distinguished from this type of distribution, which is based on actual observation, it is possible to deduce mathematically what the frequency distributions of certain populations should be. Such distributions as are expected on the basis of previous experience or probability considerations are

known as 'probability distributions' or probability distributions. For example, if a coin is tossed we expect that as n increases we shall get close to 50% heads and 50% tails. On the basis of these expectations we can test whether a given coin is unbiased or not. If a coin is tossed 100 times, we may get 44 heads and 56 tails. This is our observation. Our expectation is 50% heads and 50% tails. Now the question is whether this discrepancy is due to sampling fluctuations or is due to the fact that the coin is biased. We cannot say anything about it unless we know the expected behaviour of the coin. It should be carefully noted that the word expected or expectation is used in the sense of an average. The fact that the probabilities for both heads and tails are 1/2 does not mean that we must necessarily always get 50 per cent heads and 50 per cent tails - it only means that if the experiment is carried out a large number of times we will on an average get close to 50 per cent heads and 50 per cent tails.

A probability distribution for a discrete random variable is a mutually exclusive listing of all possible numerical outcomes for that random variable such that a particular probability distribution for the outcomes of a single roll of the die shall be as follows:

Probability distribution of the results of rolling one fair die

Face of outcome		Probability
1		1/6
2		1/6
3.		1/6
4		1/6
5		1/6
6		1/6
	Total	6/6 = 1

Since all possible outcomes are included, this listing is complete (or collectively xhaustive) and thus the probabilities must sum up to 1.

From the above table we can obtain various probabilities for the rolling of a fair die. For example,

The probability of a face = 1/6

The probability of an even face using the addition rule for mutually exclusive events is:

P(even) = P (2) + P (4) + (6)

= 1/6 + 1/6 + 1/6 = 3/6

The probability of a face of 3 or less is

P (3 or less) = P (1) + P (2) + P (3)

= 1/6 + 1/6 + 1/6 = 3/6

And the probability of a face greater than 6 is P (> 6) = 0

It should be noted that a random variable is a numerical quantity whose value is determined by the outcome of a random (chance) experiment. When a random experiment is performed, the totally of outcomes of the experiment forms a set which is a called 'sample space' (S) of the experiment. Let the random experiment be tossing of a coin 2 times. Here S = [(T.T), (T.H), (H.T), (H.H)]. If we replace T by 0 and H by 1, then the number of heads obtained in both the trials shall be:

(T.T) 0

(T.H) 1

(H.T) 1

(H.H) 2

The sample space S can be written as:

(0, 1, 2.) And here P(X = 0) = P (T.T) = 1/4

P(X = 1) = P [(T.H), (H.T)] = 1/2 P(X = 2) = P [(H.H)] = 1/4

Here P(X) = 1/4 + 1/2 + 1/4 = 1.

Such function P(X) is called the 'probability function' of the random variable X. The probability distribution is the outcome of the different probabilities taken by this function of the random variable X.

A random variable can be either discrete or continuous. A random variable is said to be discrete if the set of values defined by it over the sample space is finite. On the other hand, a random variable is

'continuous' if it can assume any (real) value in an interval. If the random variable X is a discrete one, the probability function P(X) is called 'probability mass function' and its distribution as 'discrete probability distribution' and if the random variable X is of continuous type, then the probability function P(X) is called 'probability density function' and its distribution as 'continuous probability distribution.'

Knowledge of the expected behaviour of a phenomenon or, in other words, the expected frequency distribution is of great help in a large number of problem in practical life. They serve as benchmarks against which we compare observed distributions and act as substitutes for actual distributions when the latter are costly to obtain or cannot be obtained at all. They provide decision-makers with a logical basis for making decisions and are useful in making predictions on the basis of limited information or probability considerations. For example the proprietor of a shoe store must know something about the distribution of the size of his potential customers' feet; otherwise, he may find himself with huge stock of shoes which have no market. Similarly, the manufacture of readymade garments must know the sizes of collars for which he expects maximum demand school, college or university should know what they expect of the students. It is only then that they would be in a position to comment on good or bad performance.

Amongst probability or expected frequency distributions, the following three are more popular:

1. Binomial Distribution.
2. Poisson Distribution.
3. Normal Distribution.

Among these the first two distributions are of discrete type and the last one of continuous type. It may also be pointed out that of the three distributions mentioned; the Binomial, Poisson and Normal find much wider applications in practice. Hence, they shall be discussed in detail here in this lesson.

Utility of Probability Distribution

Knowledge of the expected frequency distribution is of great help in understanding and analyzing a large number of problems of practical life. They help in taking decisions in many situations. The probability distributions are useful in the following circumstances:

1. Probability distributions serve as benchmarks against which to compare the actual frequency distributions and to find out whether the difference is due to fluctuations of sampling or some other causes.
2. They can be taken as substitutes for actual distributions when to obtain the later is costly or cannot be obtained at all.
3. Before undertaking an investigation, if a probability distribution can be constructed, it can be found out what will be the nature of distribution.
4. The probability distribution is found out under certain hypotheses or assumptions. This helps in analyzing risk and uncertainty involved in the phenomena.
5. The probability distribution helps in taking decisions on the face of uncertainty.
6. This also helps in forecasting.
7. Many of the business and other problems can be solved on the basis of probability distributions. For example, a ready-made garment manufacture decides the quantities to be produced of various sizes on the basis of 'normal distribution'. Bank-counters or railway booking where 'Q' are formed, poison distribution helps in understanding the problems and finding their solutions.

4.2 BINOMIAL DISTRIBUTION

Binomial distribution is associated with the name of James Bernoulli (1654-1705), but it was published in 1713, eight years after his death. It is also known as Bernoulli Distribution to honor its author, who did much early work on the theory and application of the binomial distribution. Binomial means two names; hence the frequency distribution falls into two categories a dichotomous process. A binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives, e.g. success or failure.

The mathematical model for a binomial distribution is developed from a very specific set of assumptions, which are

1. The number of trials is fixed.

2. There are two mutually exclusive possible outcomes on each trial which are referred to as ‘success’ and ‘failure’.
3. The probability of a success, denoted by ‘p’ remains constant from trial to trial. The probability of a failure, denoted by ‘q’ is equal to ‘1 -p2.
4. The trials are independent. That is, the outcomes on any given trial or sequence of trials do not affect the outcomes on subsequent trials.

The essential features of the binomial distribution can be stated in a general manner which enables the distribution to be used in a wide variety of circumstances. It is appropriate in any situation which involves “n” independent trials of random experiment, with two possible outcomes at each trials:

‘Success’ (with probability p) and ‘failure’ (with probability 1 - p).

It should be emphasized that the trials must all be independent and that the probability of success (and therefore of failure) does not vary between any of the trials. ‘Success’ and ‘failure’ are to be liberally interpreted as referring to the two possible outcomes, and should not be taken in their literal sense.

How binomial distribution arises, can be seen from the following examples of tossing of coins. If a fair coin is tossed once, there are two possible mutually exclusive outcomes, i.e., it may fall with head up or tail up. The probability of head coming up is: p = 1/2, and the probability of tail coming up: q = 1 - p or 1 - 1/2 = 2/3. Thus (p+q) =

1. These are the terms of the binomial (p+q) n, for n - 1.

Similarly, if two coins A and are tossed simultaneously, there are four possible outcomes:

A	**B**
H	H
H	T
T	H
T	T

(H stands for head and T stands for tail)

The probability of getting two heads simultaneously is 1 out of four outcomes or 1/4. That is also shown by our probability theorem.

In the 1st toss Probability of Coming head = 1/2

In the 2nd toss Probability of Coming head = 1/2

Probability of head coming simultaneously is 1/2 x 1/2 = 1/4. So is also true of tail.

Let p stand for the probability of success (H) and q for the probability of failure. Then the above four outcomes can be expressed as:

HH	HT	TH	TT
pp	pq	qp	qq

Or p2 2pq q^2

$p^2+2pq+q^2$ is an expansion of $(p+q)^2$.

Hence for two independent events their combined probability is given by this simple binomial expansion $(p+q)^2$.

The probability of coming head or p = 1/2

The probability of coming tail or q = 1/2 Then (p+q)2 = (1/2 +1/2)2 = 1/4+1/2+1/4.

This exactly what is explained above.

Similarly, if there are three coins A, B, C, there are following 8 possible outcomes.

A	**B**	**C**	
H	H	H	$= p^2$
H		T	
	H		

H	T	H	= $3p^2q$
T	H	H	
T	H	T	
T	T	H	= $3pq^2$
H	T	T	
T	T	T	

The probability of getting three heads is 1/8. (1/2 x 1/2 x 1/2) and that of securing two heads 3/4, and of securing one head 3/4 and securing no head 1/8. If p = 1/2and q - 1/2.

$(p+q)^3 = (1/2+1/2)3 = p^3 +3pq^2 = 1/8+3/8+3/8+1/8.$

$(a+b)^3 = a^3 +3a^2b+3ab^2 +b^3$

This expansion can further be done by adopting the following formula:

(p+q) n

If we want to know the probable frequencies of the various outcomes in a given number of trials, then the following expression will be used.

N(p+q) n

Where N stands for the number of trials. n for the number of independent events. If there are 100 trials and two independent events, the probable frequencies will be 100 $(p+q)^2$.

$= 100(p^2 +2pq+q^2)$, if p = 1/2, q = 1/2

or = 100(1/4) + 100(1/2) = 100(1/4) = 25+50+25 = 100 i.e.,

25 for two successes i.e.,

50 for one success i.e.,

25 for no success.

Similarly, if there are 100 trials and 3 independent events,

binomial expansion of N(p+q)4 = 100(1/2+1/2)3 will give the probable frequencies

$100(p^2+3p^2 q+3pq^2 +q^2)$

= 100(1/8) + 100(2/3) + 100(3/8) + 100(1/8)

= 12.5+37.5+37.5 +12.5= 100.

The formula $(p+q)^4$ for 4, 5, 6, 7, 8……n independent events will be expanded as: $(p+q)^4 = p^4 + 4p^2q + 6p^2q^2 + 4pq^2 +q^4$

$$(p+q)^4 = p^2 + 5p^4q + 10p^2q^2 + 5pq^4 +q^5$$

$$(p+q)^4 = p^3 + 6p^3q + 15p^4q^2 + 20p^3q^3 + 15p^2p^4 + 6pq^5 +q^4$$

$$(p+q)^7 = p^7 +7p^4q + 21p^5q^2 + 35p^4q^2 + 35p^2p^4 + 21p^2q^5 +q^7$$

The expansion of the formula is done in the following way:

$$(p+q)^n = pn+np^{n-1} q^1+ n(n - 1) p^{n - 2} q^2 + n(n - 1) (n - 2) p^{n - 3}q^2$$

$$1 \times 2 \qquad 1 \times 2 \times 3 \times 4 \; n(n - 1) (n - 2)$$

$$(n - 3)/ 1 \times 2 \times 3 \times 4 \times p^{n - 4}q^4 + \ldots.. + q^n$$

If n = 5 then (p+q)n = $(p+q)^5$

$(p+q)^5 = P^5+5p^{5 - 1}q^1 + 5(5 - 1)/2 \times p^{4 -2} q^2 + 5(5 - 1)(5 - 2) / 3 \times 2 \times p^{5-2} q^3 + 5(5 - 1) (5 - 2) (5 - 3)/ 4 \times 3 \times 2 \times p^{5 -4} q^4 + 5(5 - 1)(5 - 2) (5 - 3) (5 - 4) /5\times 4 \times 3 \times 2 \times 2 \times p^{5-5} q^2 = P^3+5p^4 q + 10P^3q^2+5pq^4+q^4$

The various terms of the numerical coefficients can also be obtained on the basis of laws of combination:

$$(p+q)^5 = {}^5C\, p^5q^0 + {}^5C\, p^4q^1 + {}^5C\, p^3q^2 + {}^5C\, p^2q^3 + {}^5C\, p^1q^4 + {}^5C\, p^0q^5$$

$$= 1p^5+5p^4q+ 10p^3q^2+10pq^3+5pq^4+1p^5$$

The order of p and q

The general form of binomial distribution is the expansion of (p+q)n. If this form is adopted, then the number of successes will be written in descending order. If the number of successes is to be written an ascending order, then (q +p) n will be expanded to find out the probability of each event. In such a case the expansion will be done as given below:

No of Heads (success)		$(p + q)^7$ Probability	No of Heads (success)	$(q + p)^7$ Probability
(r)	(n - r)		(r)	(n - r)
7	(0)	${}^7C_7 p^7q^0 = 1p^7$	0	7 ${}^7C_0 p^0q^7 = 1q^7$
6	(1)	${}^7C_6 p^6q^1 = 7p^6q$	1	6 ${}^7C_1 p^1q^6 = 7p^1q^6$
5	(2)	${}^7C_5 p^5q^2 = 21p^5q^2$	2	5 ${}^7C_2 p^2q^5 = 21p^2q^5$
4	(3)	${}^7C_4 p^4q^3 = 35p^4q^3$	3	4 ${}^7C_3 p^3q^4 = 35p^3q^4$
3	(4)	${}^7C_3 p^3q^4 = 35p^3q^4$	4	3 ${}^7C_4 p^4q^3 = 35p^4q^3$
2	(5)	${}^7C_2 p^2q^5 = 21p^2q^5$	5	2 ${}^7C_5 p^5q^2 = 21p^5q^2$
1	(6)	${}^7C_1 p^1q^6 = 7p1q^6$	6	1 ${}^7C_6 p^6q^1 = 7p^6q^1$
0	(7)	${}^7C_0 p^0q^7 = 1q^7$	7	0 ${}^7C_7 p^7q^0 = 1p^7$

If 8 coins are tossed continuously, then the probability of different events will be:

H	T	P	P
8	0	1/256	.004
7	1	8/256	.031
6	2	28/256	.109
5	3	56/256	.219
4	4	70/256	.273
3	5	56/256	.219
2	6	28/256	.109
1	7	8/256	.031
0	8	1/256	.004
Total		256/256	1.000 Approx.

$(q+p)^n = {}^nC\ q^n p^0 + {}^nC\ q^{n-1} p^1 + {}^nC\ q^{n-2} p^2 + {}^nC\ q^{n-3} p^3 + \ldots + {}^nC\ q^0 p^n$

$= q^n + {}^nC\ q^{n-1} p^1 + {}^nC\ q^{n-2} p^2 + {}^nC\ q^{n-3} p^3 + \ldots\ldots + p^n$

If for an event p = 1/2, q = 1/2 and n = 8, the formula will be expanded in following way:

$(1/2 + 1/2)^8 = (1/2)^8 + 8/1(1/2)^7\ (1/2) + 8(7) / (1)(2) \times (1/2)^6\ (1/2)^2$

$+ 8\ (7)(6) / (1)(2)(3) \times (1/2)^5\ (1/2)^8 + 8\ (7)(6)(5) / (1)(2)(3)(4) \times (1/2)^4\ (1/2)^4$

$+ 8\ (7)(6)(5)(4) / (1)(2)(3)(4)(5) \times (1/2)^3\ (1/2)^5 + 8\ (7)(6)(5)(4)(3) /$

$(1)(2)(3)(4)(5)(6) \times (1/2)^2 (1/2)^6$

$+ 8\ (7)(6)(5)(4)(3) / (1)(2)(3)(4)(5)(6)\ (7) \times (1/2)(1/2)^7 + (1/2)^8$

After solving it, we get:

$(1/2 +1/2)^8 = 1/256 + 8/256 + 28/256 + 56/256 +70/256 + 56/256 + 28/256 + 8/256 + 1/256$

If a production line product which is 80 percent good and 20 percent defective and a sample of size n is taken, the probability of each event in the range of the random variable can be obtained from the term resulting from raising (p+q) to the nth power. If a sample of is taken, $(P+q)^2 = p^2+pq+q^2$. The probability distribution can be obtained from the three terms in the binomial expansion:

$p^2 = .8^2$ = .64 = p (2 good, 0 bad)

$2pq = 2(.8 \times .2)$ = .32 = p (1 good, 1 bad)

$q^2 = .2^2$ = .04 = p (0 good, 2 bad)

1.00

If sample of 3 is taken, then the probability distribution is given by the expansion of

$(p+q)^3 = p^3 +3p^2q + 3pq^2 + q^3$:

$p^2 = .8^3$ = .512 = p (3 good, 0 bad)

$3p^2q = 3(.8^2 \times .2)$ = .384 = p (2 good, 1 bad)

$3pq^3 = 3(.8 \times .2^2)$ = .096 = p (1 good, 2 bad)

$q^3 = .2^3$ = .008 = p (0 good, 3 bad)

1.000

On the basis of probability the formula for binomial expansion is:

$P = n ! / r ! (n - r) ! \times p^r q^{n - r}$

4.2.1 Constants of Binomial Distribution

(a) The mean of the binomial distributing is np.

Proof. If p is the probability of success and q is the probability of failure in one trial, then in n independent trials the probabilities of 0, 1, 2, 3….., n success are given by the Ist, 2nd, 3rd …. (n + 1)th item of the binomial expansion (p+q)n. Hence, we have:

	Name of success	**Probability**	**Product**
	(x)	**(p)**	**(xp)**
	0	p^n	$0 \times q^n$
	1	$nq^{n-1}p$	$1 \times nq^{n-1}p$
	2	$n(n-1)/2 \times q^{n-2}p^2$	$2 \times n(n-1)/2 \times q^{n-2} p^2$
	….	…..	…….
	….	…..	…….
	n	p^n	$n \times p^n$
	Total	p = 1	xp

X = xp/p, p = 1

$xp = (0 \times q^n) + 1 \times nq^{n-1}p) + (2 \times n(n-1) / 2 \times q^{n-2}p^2) + \ldots + (n \times p^n)$

$= np^{n-1}p + n(n-1)\, q^{n-3}p^2 + \ldots np^n$

$= np\{q^{n-1} + (n-1)\, q^{n-2}p + \ldots p^{n-1}\}$[Taking up common]

$= np\,(q+p)^{n-1}$ [since the expansion of brackets is the expansion of the binomial $(q+p)^{n-1}$]

$= np\,(1)^{n-1} = np$

[The sum of probabilities = 1]

Thus xp = np E(X) = xp/P = np/1 = np.

Thus the mean of the binomial distribution is np.

(b) The standard deviation of Binomial Distribution is npq.

Proof: $a^2 = p_2 = v_2\ v_1{}^2$ (where v_1 and v_2 are moment about origin, zero) $v_2 = x^2p/p$

$= x^2p/1 = x^3p$

$v_2 = xp/p = xp/1 = xp = np\ a_2 = x^2p/ -$

$p\ (xp)^3$

$x^4p = 0^3q^n + 1^2.\ np^n - {}^1p^1 + 2^3\ (n - 1) / 2\ x\ q^n - {}^2\ p^2 + 3^3\ n(n - 1)\ (n - 2)/3.2\ x\ q^n - {}^3\ p^3 \ldots. + n^3p^n$

$= np^n - {}^1 + 2n(n - 1)\ q^n - {}^2p^2 + 3n(n - 1)\ (n - 2)/2\ x\ q^n - {}^3p^3 + \ldots. +\ n^3p^n$

$= np[q^n - {}^1 + 2\ (n - 1)\ q^n - {}^2p + 3(n - 1)\ (n - 2)/2\ x\ q^n - {}^2\ p^3 + \ldots. +\ n^{n-1}]$

Breaking second, third and following terms into two parts, we get:

$x2p = np[\{q^n - {}^1 + 2(n - 1)\ q^n - {}^2p + 3(n - 1)\ (n - 2)/2\ x\ q^n - {}^3p^3 + \ldots. + n^n - {}^1\} + \{$

$(n - 1)\ q^n - {}^2p + 2(n - 1)\ (n - 2)/2\ x\ q^n - {}^3\ p^3 + \ldots. + (n - 1)\ p^n - {}^1\}]$

The second term,

$2(n - 1)\ q^n - {}^3p = (n - 1)\ q^n - {}^2p^1 + q^n - {}^1p^1$

The third term

$3(n - 1)\ (n - 2) / 2\ x\ q^n - {}^3p^3 = 1.\ (n - 1)(n - 2) / q^n - {}^3p^2 + 2(n - 1)(n - 2) / 2\ q^n - {}^3p^2$

The last term,

$np^n - {}^1 = [p^n - {}^1 + (n - 1)(p^n - {}^1]$

$x^n2p = np\ [\{q+p\}^n - {}^1 + (n- 1)\ p(q^n - {}^2 + (n - 2)\ q^n - {}^3p + q^n - {}^3\}]$

$= np\ [\{(q+p)\}^n - {}^1 + (n- 1)\ p\{p+q^n - {}^2]$

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

x^3p = np [1 +(n- 1) p.1] = $(q+p)^{n-1}$ = np [1 +np - p] np + n^2p^2 - np^3.

a^2 =x^3p - $(np)^2$ = np + n^2p^2 - np^2 - (np)2.

= np + n^2p^2 - np^2 - n^2p^2 = np (1 - p) =npq [q =1 =p] a =
√npq and variance (u_3) = npq

Moments

Ist moment $u_1 = 0$

2nd moment $u_2 = a^2$ = npq

3rd moment u_3 = npq(q - p)

4th moment $u_4 = 3n^2p^2q^2 + npq(1-6pq)$

If $B_1 = 0$, the distribution will be symmetrical, i.e., there is no skewness in the distribution.

Prof. Fisher gave the following measure of skewness: $Y_1 = B_1$.

If $Y_1 = 0$, the distribution will by symmetrical, if $Y_1 > 0$, the distribution is positively skewed and if it is less than zero, the distribution is negatively skewed. V_1 can be both positive as well as negative.

$$B = U^1/U^{24} = 3n^2p^2q^2 + npq\,(1 - 6pq) / n^2p^2q^2 = 3n^2p^2q^2 / n^2p^2q^2 + (1 - 6pq) / n^2p^2q^2$$
$$= 3 + 1 - 6pq / npq$$

Prof. Fisher gave the following measure of kurtosis :$Y_2 = B_2 - 3$.

In case of distribution being normal, Y_1 will be equal to zero and Y_2 will also be zero. But converse of this is not true, that is, if Y_1 and Y_2 both are zero, the distribution may not be normal. If Y_1 is positive, the distribution is leptokurtic and Y_2 being negative, the distribution is platykurtic.

Illustration 4.1

(a) What is the probability of obtaining three aces in rolling a die three times?

(b) The administrator of a large airport is interested in the number of aircraft departure delays that are attributable to inadequate control facilities. A random sample of 10 aircraft takes off is to be thoroughly investigated. If the true proportion of such delays in all departures is .40, what is the probability that 4 of the sample departures are delayed because of control inadequacies?

Solution:

(a) Probability of getting an ace in a single throw of dice = 1/6.

'p' of getting three aces in rolling a die three times.

= 1/6 x 1/6 x 1/6 = 1/216.

According to the formula:

$p = n! / r! (n - r)! p^r q^{n - r}$

$= [3 ! / 3 ! (3 - 3) !] \times [1/6]^3 \times [5/6]^0 = 216$

(b) Let the sample investigations be considered as trials in a Binomial distribution, and a control caused delays a success (p) = .40.

$p (4) = n ! / r ! (n - r) ! p^r q^{n - r}$

$= 10 ! / 4 ! (10 - 4) ! \times (.4)^4 \times (1 - .4)^{10 - 4}$

$= 210 (.4)^4 (.6)^6$

$= 210(.256)(.046656) = .2508$

Illustration 4.2

A bowl contains 4 red balls and 6 green balls. Five balls are to be drawn with replacement. Find out probabilities for different events.

Solution:

The probability of drawing a red ball = 4/10 = .4(p)

The probability of drawing a red ball = 6/10 = .6(p)

Formula;

$p = n ! / r ! (n - r) ! \times p^r q^{n-r}$

Number of balls to be drawn		Probability
Red	Green	
5	0	$p = 5 ! / 5 ! (5 - 5) ! \times (.4)^5 (.6)^0 = .01024$
4	1	$p = 5 ! / 4 ! (5 - 4) ! \times (.4)^4 (.6)^1 = .07680$
3	2	$p = 5 ! / 3 ! (5 - 3) ! \times (.4)^3 (.6)^2 = .23040$
2	3	$p = 5 ! / 2 ! (5 - 2) ! \times (.4)^2 (.6)^3 = .34560$
1	4	$p = 5 ! / 1 ! (5 - 1) ! \times (.4)^1 (.6)^4 = .25920$
0	5	$p = 5 ! / 0 ! (5 - 0) ! \times (.4)0 (.6)^5 = .07776$
		1.00000

The same results are obtained by other formula also. $(p+q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$

Number of ball to be drawn			
Red	Green	Computation of Probability	Probability
5	0	$p = p^5 = (.4)^5$	= .01024
4	1	$p = 5p^4q = 5(.4)^4(.6) = 5 (0256)(.6)$	= .07680
3	2	$p = 10p^3q^2 = 10(.4)^3(.6)^2$ $=10(.064)(.36)$	= .23040
2	3	$p =10p^3q^2 = 10(.4)^2(.6)^3 =$ $10(.16)(.216)$	= .34560

1	4	$p = 5pq^4 = 5(.4)(.6)^4 = 5\ (.4)(.1296)$	$= .25920$
0	5	$p = q^5 = (.6)^3$	$= .07776$
		Sum of the probabilities	1.00000

Illustration 4.3

(a) Assuming that the half population are vegetarians so that the chance of an individual being a vegetarian is 1/2, and assuming that 100 investigation can take a sample of 10 individuals to see whether they are vegetarians, how many investigators would you expect to report that three people or less were vegetarians.

(b) The incidence of occupational disease in an industry is such that the workers have a 20% chance of suffering from it. What is the probability that out of six workers 4 or more will contact disease?

(c) A company makes TVs, of which 15 per cent are defective. Fifteen TVs are shipped to a dealer. If such TV assembled is considered an independent trial, what is the probability that the shipment of 15 TVs contains (i) no defective TV, (ii) one or less defective YV.

Solution:

(a) Probability of a person being vegetarian (p) = .

1/2 q = 1 - p = 1- 1/2 - 1/2.

Out of 10, 3 or less vegetarian

3	7	$p = 10\ !\ /\ 3\ !\ (10 -3)\ !\ (.5)^3(.5)^7$	$=.1171875$
2	8	$p = 10\ !\ /\ 2\ !\ (10 -2)\ !\ (.5)^2(.5)^8$	$=.0439453125$
1	9	$p = 10\ !\ /\ 1\ !\ (10 -1)\ !\ (.5)^1(.5)^9$	$=.00365625$
0	10	$p = 10\ !\ /\ 0\ !\ (10 -10)\ !$ $(.5)^0(.5)^{10}$	$=.0009765625$

0 .1718750000

For 100 investigators the expected number np = 100 x .171875 = 17.1875

(b) The probability of a man suffering from disease or p = 20/100 = .2 q = 1 - p = 1- .2 = .8

	Contacting disease	Not contacting disease	
4	2	$p = 6 ! / 4 ! (6 -4) ! (.2)^4(.8)^2$	=.01536
5	1	$p = 6 ! / 5 ! (6 -5) ! (.2)^5(.8)^1$	=.001536
6	0	$p = 6 ! / 6 ! (6 - 6) ! (.2)^6(.8)^0$	=.000064
			0.016960

p = .016960.

(c) p = 0.15, q = 1 - p = 1 - 0 .15= 0.85.

p (No defective TV) = $n ! /r ! (n - r) ! \times p^n q^{n - r} = 15 ! / 0 ! (15 -0) ! \times (.15)^5(.85)^{10}$

$=(.85) =.0873$

p (1 defective TV) = $15 ! / 1! (15 - 1) ! \times (.15)^1(.85)^{14} =15 (.15)^1 (.85)^{14}$ .2312.

Probability of one or less defective TV = .0873 + .2312 = .3185.

Illustration 4.4

Six dice were thrown 128 times. Each 4, 5 or 6 spot appearing was considered to be success and each 1, 2, or 3 spot a failure. The results were:

No. of Success	0	1	2	3	4	5	6
	0	10	26	44	34	14	0

Compare actual and probability frequencies. Calculate means and standard deviations of actual and probability distribution also.

Solution:

Expected frequency = $N(p+q)^6$

= $128(p^6+6p^5q+15p^4q^2+20p^3q^3+15p^2q^4+5pq^5+q^6)$

= 128(2 / 64) +128(4 / 64) +128(15 / 64) +128(20 /64) +128(15 / 64) +128(4 / 64) +128(1 / 64).

	= 2, 12, 30, 40, 30, 12, 2.	
No of Successes	**Actual Frequencies**	**Expected Frequencies**
0	0	2
1	10	12
2	26	30
3	44	40
4	34	30
5	14	12
6	0	2
	128	128

Mean and Standard Deviation of Actual and Probability Frequencies

X	**f(A)**	**fX(A)**	**d,=(X-3)**	**fD,(A)**	**fD^2(A)**	**f(E)**	**fX(E)**	**fd,(E)**	**fd^2(E)**
0	0	0	-3	0	0	2	0	-6	18
1	10	10	-2	-20	-40	12	12	-24	48
2	26	52	-1	-26	26	30	60	-30	30
3	44	132	0	0	0	40	120	0	0
4	34	136	+1	+34	34	30	120	+30	30

5	14	70	+2	+28	56	12	60	+24	48
6	0	0	+3	0	0	2	12	+6	18
	128	**400**		**+ 16**	**156**	**128**	**384**		**192**

Actual Frequencies Probability Frequencies

X = fX (A) / N = 400/128 = 3.125, X = fX (E) / N = 384/128 = 3

a = fd ^{2} / N ⇒ (fd / N)^3$_2$ a = fd ^{2} / N = 192 / 129

= 156/128 - (16 /128)^2 = 1.1 = 1.5 = 1.22

The mean and S.D. of the probability frequencies can also be found out by the following formulas:

X = Mean, n = number of independent events, p = probability of success X = 6 x 1/2 = 3. Standard Deviation:

a = npq = 6 x 1/2 x 1/2 = 1.5 = 1.22 (the same calculated above)

Illustration 4.5

(a) The probability of the birth of a male child is 1/2. 4096 families were chosen at random having just four children. Find out the probabilities of having 0, 1, 2, 3 and 4 male children in a family and ascertain the probability frequency distribution based on those probabilities. Find the mean and standard deviation of the distribution.

If the probability of the birth of a male child is 1/4, what will be the probabilities of having 0, 1, 2, 3 and 4 male children in a family?

(b) If the 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 two bolts will be defective.

(c) The following data shows the number of seeds germinating out of 10 on damp filter for 80 sets of seeds. Fit a binomial distribution to the data:

X:	0	1	2	3	4	5	6	7	8	9	10
f:	6	20	28	12	8	6	0	0	0	0	0

Solution:

(a) The probability of the birth of a male child: p = 1/2 q = 1 - p = 1 - 1/2 = 1/2

$$N(p+q)^n = 4096\,[1/2 + 1/2]^4$$

Binomial distribution terms: $4096(p^4+4p^3q+6p^2q^2+4pq^3+q^4)$

$= 4096\{[1/2]^4 + 4[1/2]^3\,[1/2] + 6[1/2]^2\,[1/2]^2 + 4[1/2][1/2]^3 + [1/2]^4\}$

$= 4096\,\{1/16 + 4/16 + 16/16 + 4/16 + 1/16\}$

= 256, 1024, 1536, 1024, 256.

X = np = 4 x 1/2 = 2.

a = npq = 4 x 1/2 x 1/2 = 1.

If the probability of the birth of male child is : p = 1/4. q = 1 - q = 1 - 1/4 = 3/4

$N(p+q)^n$, $4096\,[1/4 + 3/4]^4$.

$= 4096\,\{[1/4]^4 + 4[1/4]^3\,[3/4] + 6[1/4]^2\,[3/4]^2 + 4[1/4][3/4]^3 + [3/4]^4\}$

$= 4096\,\{1/256 + 3/64 + 27/128 + 27/64 + 81/256\}$

= 16, 192, 864, 1728, 1296.

(b) Probability that out of 4 bolts chosen at random:

(i) I will be defective:

$p = n! / r!(n - r)! \times p^r q^{n-r}$

$= 4! / 1!(4 - 1)! \times (.2)(.8)^2$

(ii) I will be defective:

$p = n! / r!(n - r)! \times p^r q^{n - r}$

$= 4! / 0!(4 - 0)! \times p^0 q^4$

$= (.8)^4 = .4096$

(iii) 2 will be defective:

$p = n! / r!(n - r)! \times p^r q^{n - r}$

$= 4! / 2!(4 - 2)! \times p^2 q^2$

$= 6(.2)^2 \quad (.8)^2 = .1536$

The probability that at most two bolts will be defective is

=.4096+.4096+.1536 = .9728

(c)

X:	0	1	2	3	4	5	6	7	8	9	10
f:	6	20	28	12	8	6	0	0	0	0	0
fX:	0	20	56	36	32	30	0	0	0	0	0

$\bar{X} = 174/80 = 2.175$

$X = np = 174/80$

$p = 174/80 \times 10 = 174/800 = .2175$ $q = 1 - p = 1 - .2175 = .7825.$

The binomial distribution to be fitted to the data: $N(q+p)^n = 80 (.7825 + .2175)^{10}$

X (r)	(n - r)		Probability		Frequencies
0	10	$^{10}C_0p^0q^{10} = q^{10}$	$80(.7825)^{10}$	=	6.9
1	9	$^{10}C_1p^1q^9 = 10p^1q^9$	$80(10)(.7825)^9(.2175)$	=	19.1
2	8	$^{10}C_2p^2q^8 = 45p^2q^8$	$80(45)(.7825)^8(.2175)^2$	=	24.0
3	7	$^{10}C_3p^3q^7 = 120p^3q^7$	$80(120)(.7825)^7(.2175)^3$	=	17.8
4	6	$^{10}C_4p^4q^6 = 210p^4q^6$	$80(210)(.7825)^6(.2175)^4$	=	8.6
5	5	$^{10}C_5p^5q^5 = 252p^5q^5$	$80(252)(.7825)^5(.2175)^5$	=	2.9
6	4	$^{10}C_6p^6q^4 = 210p^6q^4$	$80(210)(.7825)^4(.2175)^6$	=	0.7
7	3	$^{10}C_7p^7q^3 = 120p^7q^3$	$80(120)(.7825)^3(.2175)^7$	=	0.1
8	2	$^{10}C_8p^8q^2 = 45p^8q^2$	$80(45)(.7825)^2(.2175)^8$	=	0.0
9	1	$^{10}C_9p^9q^1 = 10p^9q^1$	$80(10)(.7825)^1(.2175)^9$	=	0.0
10	0	$^{10}C_{10}p^{10}q^0 = p^{10}$	$80(.2175)^{10}$	=	0.0
					80.1

4.2.2 Properties of the Binomial Distribution

(1) The shape and location of binomial distribution changes as p changes for given n. Suppose, when p = .1, the probability distribution is quite positively skewed, with the probabilities of small values being greatest. As p becomes larger, the skew becomes less pronounced. When p = .5, the distribution becomes symmetrical, so that the probability of particular value is equal to the probability of 1 - p. When p is larger than .5, the distribution becomes negatively skewed. Notice that the probabilities for p = .1 are identical, but in reverse sequence, to chart for p = .9. This will hold for every complementary pair of p values, such as, .3 and .7, .01 and .99 etc.

(2) The mode of the binomial distribution is to the value of x which has the largest probability. If n = 6, and p = .3, the mode is 2. If n = 6 and p = .9, the mode will be 6. The mean (X) and mode (M0) are equal if np is an integer. If n = 6, and p = .5, the mean and mode will be equal and their value will be 3. For fixed n, both mean and mode will increase with the increase in p.

(3) The mean of the binomial distribution thus increases obviously as n increases with p remaining constant.

(4) If n is large and if neither p nor q is too close to zero, the binomial distribution may be approximately normal distribution with standardized variable given by Z

= [X – np] / npq.

4.2.3 Importance of Binomial Distribution

The binomial distribution is often very useful in decision-making situation in business. One area in which it has been very widely applied is quality control. In acceptance sampling plans, inspection is carried out on the articles drawn in a sample, and lots or shipments are either accepted or rejected on the basis of sample evidence. Such plans are widely used in industry for incoming materials, at various stages of manufacturing processes, for outgoing final product, and for inspection by purchases of material supplied by suppliers.

The binomial probability distribution is a discrete probability distribution which describes an enormous variety of real life events. However, it is important before using the distribution that there is correspondence between the real life situation and the model. In many cases the underlying assumptions of binomial distribution are obviously not met. For example, suppose that in a production process items produced by a certain machine tool are tested as to whether they meet specifications. If the items are tested in the order in which they are produced, then the assumption of independence would doubtless be violated. That is, whether an item meets specifications would not be independent of whether the preceding item (s) did. If the machine tool had become subject to wear, it is quite likely that if it produced an item which did not meet specifications, the next item would fail to conform to specifications in a similar way. Thus, whether or not an item is defective would depend on the characteristics of preceding items. On the other hand, in coin-tossing, an experiment getting head or tail on a particular toss does not affect the outcome on the next toss.

It is seen from the assumptions underlying the binomial distribution that it is applicable to the situations of sampling from a finite universe, with replacement or sampling from an infinite universe, with or without replacement. In either of these cases, the probability of success may be viewed as remaining constant from trial to trial and the outcomes as independent among trials. If the population size is large relative to sample size, this is, if the sample constitutes only a small fraction of the population, and if p is neither very close to zero or one in value, the binomial distribution if often sufficiently accurate, even though sampling may be carried out from a finite universe without replacement. Furthermore, in general, approximations are relatively closer for terms near the center of the distribution than in the trials and for sum of the terms rather than for individual terms.

4.3 POISSON DISTRIBUTION

Simeon Danis Poisson (1781-1840), a noted French mathematician of his time developed a concept of frequency distribution in 1837, who is said to have been studying the number of persons kicked to death by mules in various army divisions employing mules for transportation of equipment and personnel. It is another very useful probability distribution named after his name. The Poisson distribution is based on the same assumptions as the binomial distribution. This means that in a Poisson experiment we deal with either success or failure, that the successes are independent of each other, and thus, the probability of success throughout the entire process remains constant. However, the Poisson distribution can be viewed as a limiting form of binomial distribution when n approaches infinity (n - 00) and p approached zero (p 0) in such a way that their product is some fixed number (m), i.e., it remains constant. In other words, Poisson distribution is applicable where there are a number of random situations where the probability of a success on a single trial is small and the number of trials is large. It is used as a model to describe the probability distribution of such events as the insurance claims, breakdowns of machines, number of typing errors per page, arrival of customers etc. Events that are generated by a Poisson process require a given interval (of time, length, or space) that can be divided into small intervals, each of which can be considered as a trial. The sub-intervals should be sufficiently small so that no more than one success can occur in a sub-interval. The number of sub-intervals will be the number of trials. The trials are assumed to be independent. For example, in the queuing theory (waiting line), one may assume that the arrivals coming for service per unit of time constitute a Poisson process. Computing probabilities by direct use of the binomial distribution for a large number of trials is a long and tedious

task. On the other hand, the Poisson distribution with less demanding computations can be used as an approximation to the binomial distribution. Poisson distribution is derived as an approximation of the binomial distribution, rather than from the Poisson process.

In case of binomial distribution the probability of r successes is given by:

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

$n(n-1)(n-2)\ldots(n-r+1)/r!\ x\ p^{r}q^{n-r}$

Put $p = m/n$ $q = 1 - p = 1 - m/n$

Thus we get

$p(r) = n(n-1)(n-2)\ldots(n-r+1)/r!\ x\ (m/n)^{r}(1-m/n)^{n-r}$

$1(1-1/n)(1-2/n)\ldots(1-r-1/n)m(1-m/n)^{n}/r!\ (1-m/n)^{r}$

For fixed r_2 as n 00

$[1-1/n][1-2/n]\ldots[1-r-1/n][1-m/n]^{r}$ all tend to 1 and $[1-m/n]^{n}$ to e^{-m}.

Thus in the limiting case $p(r) = e^{-m}m^{r}/r!$ or $e^{-m}\ x\ m^{r}/r!$

This is called the Poisson probability distribution.

Here,

e = the base of the natural logarithms and has a value of 2.7183.

m = positive constant equal to the mean of the distribution.

r = any positive integer for which probable frequency is to be calculated.

The Poisson distribution is a discrete distribution with a single parameter m. With the increase in the value or m, the distribution shifts to the right.

4.3.1 Form of Poisson Distribution

Just like binomial distribution the variate of the Poisson distribution is also discrete one, that is, it takes only integral values. The probabilities of 0, 1, 2, 3….. Successes can be found out by successive terms of the expansion:

$$p = e^{-m} [1 + m /1! + m^2 /2! + m^3 /3! + m^4 /4! + …… + m^r /r! +….]$$

It can also be written in the following form

No of successes: 0	2	3	4	….. r
Probabilities (p): e^{-m}	$me^{-m}/2!$	$m^3e^{-m}/3!$	$m^4e^{-m}/4!$	$….m^re^{-m}/r!$

The Calculating Process

To fit a Poisson distribution, the probabilities of 0, 1, 2, 3, 4 …. are found out, as discussed below:

(i) First of all the arithmetic mean (m) of the data is calculated.

(ii) The value of e^{-m} is obtained.

The value of e is 2.7183 (the base of natural logarithm).

$e^{-m} = 1/e^m = 1/(2.7183)^m$ = 1/Antilog (Log 2.7183 x m)

=1/Antilog (.4343 x m) = Reciprocal of [Antilog (.4343 x m)]

e^{-m} = Reciprocal of [Antilog (.4343 x m)]

(iii) By using the following formula, the probabilities of 0, 1, 2, 3, 4….. successes according to Poisson distribution will be obtained.

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

The probabilities may also be obtained using the relation. $P(r) = m^r$

$^{-1}e^{-m}$ /(r - 1) ! m/r = m/r x P(r - 1)

(iv) Finally expected frequencies are obtained as

$$N e^{-m} x m^r/r!$$

Where N is the total observed frequency.

Illustration 4.6

Fit a Poisson's distribution to the set of the observations:

Deaths	0	1	2	3	4
Frequency	122	60	15	2	1

and calculate the probability frequencies.

Solution

X	f	fX
0	122	0
1	60	60
2	15	30
3	2	6
4	1	4
	200	**100**

M = fX / N = 100/200 = 0.5

The probability for 0 deaths

$$P_{0)} = e^{-m} x m^0/0!$$

= Reciprocal [AL (Log 2.7183 x .5)]

= Reciprocal [AL (.4343 x .5)]

= Reciprocal [(AL of .21715)

= Reciprocal of 1.649 = .6065

The expected number of 0 deaths in 200 cases = .6065 x 200 = 121.30

$P_{(1)} = e^{-m} \times m_1/1!$ or $P_{(0)} \times m = 121.30 \times .5 = 60.65$
$P_{(2)} = e^{-m} \times m_2/2!$ or $P_{(1)} \times m/2 = 60.65 \times .5/2 = 15.16$
$P_{(3)} = e^{-m} \times m_3/3!$ or $P_{(2)} \times m/3 = 15.16 \times .5/3 = 2.52$

$P_{(4)} = e^{-m} \times m_4/4!$ or $P_{(3)} \times m/4 = 2.52 \times .5/4 = 0.94$

Hence, expected frequency distribution will be:

X	0	1	2	3	4	Total
f	121	61	15	2	1	200

Note: $m^0/0!$ is taken as equal to one and not equal to zero.

The value of e^{-m} can also be obtained by referring to the following table in the manner explained below:

Table of Values of e^{-m}

(m is lying between 0 and + 1)

m	0	1	2	3	4	5	6	7	8	9
0.0	1.000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	.9048	.8958	.8869	.8781	.8694	.8607	.8521	.8437	.8353	.8208
0.2	.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	.5488	.5434	.5379	.5326	.5273	.5220	.5169	.5117	.5066	.5016
0.7	.4966	.4916	.4868	.4819	.4771	.4724	.4677	.4630	.4584	.4638
0.8	.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107

0.9	.4066	.4025	.3985	.3946	.3906	.3867	.3839	.3791	.3753 .3716

For example, $e^{-m} = .6065$. The same value was obtained in the previous example. Similarly, $e^{-57} = 0.5655$.

Table of e^{-m}

(m is lying between +1 and + 10)

m	1	2	3	4	5
e^{-m}	0.36788	0.13534	0.04979	0.01832	0.006733
m	6	7	8	9	10
e^{-m}	0.002479	0.000912	0.000335	0.000123	0.000045

Example $e^{-2.57} = e^{-2.0} \times e^{-.57} = 0.13534 \times 0.5655 = 0.07653$

$e^{-1.5} = e^{-1} \times e^{-.5} = .36788 \times .6065 = .22312922$ or $.2231$

Illustration 4.7

In 1000 consecutive issues of the Utopian Seven Daily Chronicle the deaths of centenarians were recorded, the number x having frequency 'f' according to the table.

x :	0	1	2	3	4	5	6	7	8
f :	229	325	257	119	50	17	2	1	0

Show that the distribution is roughly Poissonian by calculating its mean and then the frequencies in the Poissonian distribution with the same mean and the same total frequency of 1000. (Given $e^{-1} = .2231$ approx.)

Solution:

The number of cases with 0 deaths, in 1000.

$p_{(0)} = e^{-m} \times m^0/0 ! \times 1000 = .2231 \times 1000 = 223.1$

Similarly:

$p_{(1)} = e^{-m} \times m^1/1 ! \times 1000 = .2231 \times 1.5 \times 1000 / 1 = 334.65$

$p_{(2)} = e^{-m} \times m^2/2 ! \times 1000 = .2231 \times 1.5 \times 1.5 \times 1000 / 2 \times 1 = 25098.75$

$p_{(3)} = e^{-m} \times m^3/3 ! \times 1000 = 125.49375$
$p_{(4)} = e^{-m} \times m^4/4 ! \times 1000 = 47.60156$
$p_{(5)} = e^{-m} \times m^5/5 ! \times 1000 = 14.280468$
$p_{(6)} = e^{-m} \times m^6/6 ! \times 1000 = 3.570117$
$p_{(7)} = e^{-m} \times m^7/7 ! \times 1000 = .7140234$
$p_{(8)} = e^{-m} \times m^8/8 ! \times 1000 = .14280468$

Following are the observed and the expected number of cases:

X	Observed Frequency	Expected Frequency
0	229	223.10
1	325	334.65
2	257	250.99
3	119	125.49
4	50	47.60
5	17	14.28
6	2	3.57
7	1	0.71
8	0	0.14
Total	1000	1000.53

Note: The difference of .53 is on account of approximation.

4.3.2 Constants of the Poisson Distribution

The constants of the Poisson distribution can be obtained by putting 1 in place of q as value of q is almost equal to 1.

(i) Mean:

X or m = np = n x X/n = X or m [p = X/n]

(ii) Standard Deviation:

σ = n x p x q = n x p x 1 = np = m or σ^2.

Thus, in a Poisson distribution

Mean = Variance = σ^2

Proof:

Poisson Distribution is given as under:

x	p(x)	x, p(x)	x^2, p(x)
0	e^{-m}	0	0
1	e^{-m}, m	e^{-m}, m	e^{-m}, m
2	e^{-m}, m^2/2 !	e^{-m}, m^2/2 x 2	e^{-m}, m^2/4 x 2
3	e^{-m}, m^3/3 !	e^{-m}, m^3/3 x 2 x 3	e^{-m}, m^3/3 x 2 x 9
3	e^{-m}, m^4/4 !	e^{-m}, m^4/4 x 3 x 2 x 4 e^{-m},	m^4/4 x 3 x 2 x 16
Total	p(x)	xp (x)	x^2, p(x)

X or m = x p(x)/ p(x) σ = x^2, p(x)/p(x) - (x, p/p)^2

$p(x) = e^{-m} [1 + m + m^2/2! + m^3/3! + m^4/4! \ldots]$

But according to exponential theorem:

$1 + m + m^2/2! + m^3/3! + m^4/4! \ldots e^m$

$p(x) = e^{-m} x^* e^m = 1/e^m \times e^m = 1$ x.

$p(x) = 0 + e^{-m}, m + e^{-m}, m^2 + e^{-m}, m^3/2 + e^{-m}, m^4/6 + \ldots$

$= e^{-m}, m [1 + m + m^2/2! + m^3/3! + \ldots.]$

$= e^{-m}, m \times e^m = e^{-m} \times e^m = 1/e^m \times e^m \times m = n$

$X = x. p(x) / p(x) = m/1$ Hence $m = np$.

$x^2. p(x) = 0 + e^{-m}, m + e^{-m}, m^2 / 1! \times e^{-m}, m^3/3! + e^{-m}, m^4/4! \times 4^3 \ldots$

$= e^{-m}, m [1+ (m/2 \times 4) +(m^2/ 3 \times 2 \times 9) +(m^3/ 4 \times 3 \times 2 \times 16) + \ldots]$

$= e^{-m}, m [1+ 2m + 3. m^2/ 2 ! + 4 \times m^3/ 3 ! \ldots]$

Breaking the terms within the brackets into two parts:

$x^2. p (x) = e^{-m}, m [\{1+ m + 1\ m^2/ 2 ! + 1m^3/ 3 ! + \ldots\} + \{m + 2m^2/2 ! + 3m^3/ 3! + \ldots\}]$

$= e^{-m}, m [e^m + m\{1 + m + m^2/ 2 ! + m^3/ 3 !\}]$

$= e^{-m}, (e^m + e^m) = e^{-m}, m. e^m (1+ m)$

$= 1/ e^m \times e^m \times m(1 + m) = m. e^m (1+ m) = m + m^2$

$\sigma = x^2. p(x) / p(x) - [xp(x) / p(x)]^2 = m + m^2 / 1 - [m/1]^3$

$\sigma^2 = u_2 = m, q = m = np$

The four moments in a Poisson distribution will be:

First Moment $u_1 = 0$

Second Moment $u_2 = m$

Third Moment $u_3 = m$

Fourth Moment $u_4 = m+3m^3 = m\ (1+ 3m)$

$$B_1 = u_3^2/u_2^3 = m^2/m^3 = 1/m$$
$$B_4 = u_4/u_2^2 = 3m^2 + m\ /m^2 = m(3m + 1)m.\ m = 3m + 1/m + 1/m = 3 + 1/m.$$

In a Poisson distribution, if only mean (m) is known, all other constants can easily be computed. This is one of the great advantages of Poisson distribution.

4.3.3 Characteristics of Poisson Distribution

The Poisson distribution possesses the following characteristics:

(1) Discrete Distribution. Like binomial distribution, Poisson distribution is also a discrete distribution, that is, it is concerned with occurrences that can be described by a discrete random variable, denoting number of successes by 0, 1, 2, 3… However, a Poisson random variable can assume any one of infinite number of values, 0, 1, 2…, whereas a binomial variable is limited to the values 0, 1, 2, …..n.

(2) The values of p and q. It is applied in conditions where the probability of the success of an event (p) is very small and that of failure (q) is very high, almost equal to 1, and n is also very large.

(3) Main Parameter. The main parameter of the Poisson distribution is mean (m = np). If the value of m is known, the values of other parameters can be ascertained very easily.

(4) Values of Constants. The following are the values of constants in the Poisson distribution.

X or m = np, a= m = np.

$\mu_1 = 0$, $\mu_2 = m$, $\mu_3 = m$, $\mu_4 = m\ (3m +1)$.

(5) Form of Distribution. The Poisson distribution is the skewed distribution. With the increase in the value of m, the distribution shifts to right and skewness is reduced.

(6) Assumptions. The Poisson distribution is based on the following assumptions; (i) the occurrence or non-occurrence of an event does not influence the occurrence or non-occurrence of any other; (ii) the probability of a success for a small time interval or region of space is proportional to the length of the time interval or region of space; (iii) the probability of more than one event happening in a very small interval is negligible.

(7) Utility. The Poisson distribution fits a very good model for use for determining probabilities associated with random variables where p is very small and n is very large, such as, the number of calls coming into a telephone switch-board, the number of defects in manufactured part, number of accidents, number of customers arriving at a service facility, etc.

4.4 NORMAL DISTRIBUTION

The binomial and Poisson distribution are the most useful probability distributions of discrete random variables. Probability or probability distributions of continuous random variables are also of considerable importance in statistics. The normal distribution, also called the normal probability distribution, is the most useful probability distribution for continuous variables. The normal distribution is perhaps the most important distribution encountered in statistical applications. One very good reason for this is that so many physical measurements and natural phenomena have actual observed frequency distributions that closely resemble the normal distribution. In fact, the normal distribution plays a central role in statistical theory and practice, particularly in the area of statistical reference, and can be regarded as the cornerstone of modern statistics.

The normal distribution was first discovered by the English mathematician De Moiver (1667-1754). It was later rediscovered and applied in sciences, both natural and social, and in practical affairs by the French Mathematician Laplace (1749-1827). It was extensively developed and utilized by the German Mathematician, physicist and astronomer, Carl Gauss (1777-1855), who was one of the first persons to describe its mathematical properties. The normal distribution is sometimes also called in honour of Gauss as Gaussian distribution. One of the first to make an extensive use of the normal distribution in social statistics was the Belgian astronomer and statistician Quetelet (1796-1874). A pioneer in its application to biological data was the English anthropologist, biometrician, criminologist, geneticist, meteorologist, psychologist and statistician, Sir Francis Galton (1822-19110, a cousin of Charles Darwin.

The normal distribution is the most frequency used of all probability distributions. The probability distributions of most sample statistics are derived and closely connected with normal distribution. The fundamental importance of the normal distribution in statistics arises from the fact that the distribution of sample means and many other statistics for large sample sizes is approximately normal, even though the original population may not be normal. If the population from which samples are drawn is normally distributed, the means of the samples are normally distributed around the time mean regardless of the size of the sample. The normal distribution has convenient mathematical properties and it also serves as an approximation to other discrete probability distributions, such as binomial, Poisson, etc. It is particularly useful for approximating the binomial distribution, if the sample size is large because the binomial distribution is unimodal, it is symmetrical if p = .5, and it approached the normal distribution as the sample size approaches infinity. It is very close to normal distribution and may be successfully approximated even when p is not equal to .5, if n is large. The normal distribution can be used to approximate Poisson distribution, if the mean is large.

4.4.1 Importance of Normal Distribution

If a statistician could select but one distribution to work with during his life time, he would almost surely select the normal distribution. Although modern statisticians and applied economists could perhaps get along without computer, it would be exceedingly difficult to do without the normal distribution. This statement indicates the importance of normal distribution in the theory of statistics. The following points will highlight the importance of normal distribution:

1. The normal distribution has the property stated in the central limit theorem, which has made it of such fundamental importance in statistics. The central limit theorem states that, ‘as the sample size n becomes large, when each of the observation is independently selected from a population having a mean of u and a standard deviation of a, the sampling distribution of X tends towards a normal distribution with a mean of u and a standard deviation of a/n. The central limit theorem is applicable regardless of the shape of the population frequency distribution. It is valid for populations having the skewed, binomial, uniform or exponential frequency distributions. It may be used whether the observed random variable is discrete or continuous. In each case, as the sample size increase, the sampling distribution of X becomes normal. This characteristic makes it possible to determine the minimum and maximum

limits within which the population values lie. For example, within a range of population mean = 3a 99.73% (normally all) items are covered.

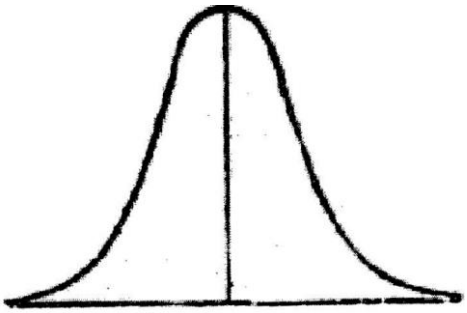
2. As n becomes large, the normal distribution gives a good approximation of many discrete distributions, such as binomial, Poisson, etc., where exact discrete probability is difficult or impossible to obtain correctly.

3. In probability as well as in applied statistics, there are many problems which can be solved under the assumption of normal distribution with satisfactory results.

4. The mathematical properties of the normal distribution make it popular and comparatively easy to manipulate. Many of the statistical techniques of describing and interpreting data are directly attributable to the properties of the normal curve. Without it the technique of sampling could not have developed.

4.4.2 The Shape of the Normal Curve

The normal curve is bell shaped, symmetrical and asymptotic in both directions to the x-axis and depends on the two parameters, X and only. That is, we need to know only the mean and standard deviation to be able to compute the entire distribution. While it is always bell-shaped and symmetrical about its mean, its actual shape is determined by the standard deviation of the distribution. The following figure indicates this:



The normal curve is represented in several forms. The following is the basic equation of the normal curve:

$$P(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\bar{X}}{\sigma}\right)^2}$$

Where,

y = The computed height of an ordinate at a distance of π from the mean.

σ = Standard deviation of the given normal distribution.

$\bar{X}$ = The constant, 22/7, or 3.14172

e = The constant 2.7183 (The base of Natural Logarithms).

With the help of the above formula, the desired probability is found out.

4.4.3 Properties of the Normal Curve

The normal curve has the following properties:

• It is bell-shaped and symmetrical curve.

• The normal curve is unimodal, symmetrical about its peak and its tails extend indefinitely in both directions, that is although the curve comes closer and closer to the horizontal axis, it never touches it.

• All its measures of central tendency are equal (Mean = Median = Mode), and they are located at the abscissa of the highest ordinate.

• All observations are included inside the curve above the x-axis.

• The mean ordinate cuts the curve into two equal parts. The scatter of the frequencies on the one side is exactly the same as it is on the other side.

• Near the mean value, the normal curve is concave while near # 3a it is convex to the horizontal axis. The points of inflection, i.e., the points where the change in curvature occurs are # 1a.

• Within a range .6745 of the a on both sides of the mean 50% of the frequencies occur. This is probable error.

• First and Third quartiles are at equal distance from the median. (Q_3 = M) = (M – Q_1).

• Quartile Deviation = Probable Error.

• Mean Deviation is .7979 or 4/5 of the Standard Deviation. If mean deviation is added to the lower or subtracted from the upper quartile gives the Mean = Median = Mode.

• The probable error is .845 of Mean Deviation.

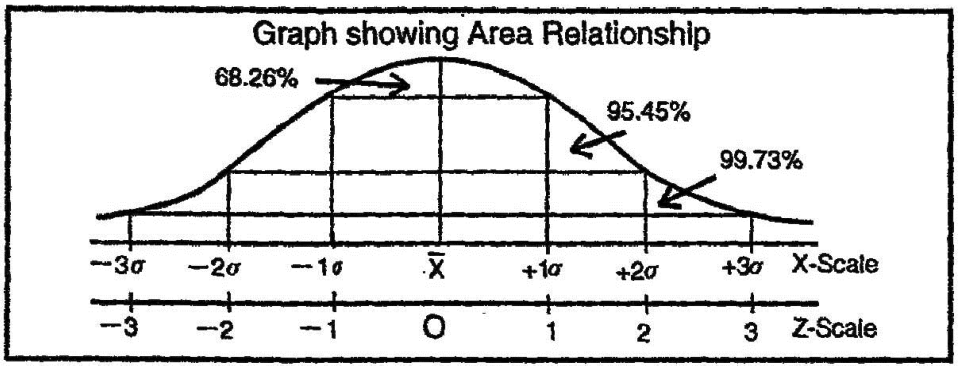
• The total area under the normal curve is 1. Its area relationship is as follows:

(i) Mean ± 1□ covers 68.26% area. 34.134% area will lie on either side of the mean.

(ii) Mean ± 2□ covers 95.45% area. 47.725% area will lie on either side of the mean.

(iii) Mean ± 3□ covers 99.73% area. 49.865% area will lie on either side of the mean.

The following figure illustrates the area property



4.4.4 Conditions of Normality

The following four conditions must prevail among the factors affecting individual events that make up a given population, if the distribution is to be normal:

1. The causal forces are numerous and of approximately equal weight.

2. These forces must be the same over the universe from which the observations are drawn. Though their incidence will vary from event to event. This states the condition of homogeneity.

3. The forces affecting events must be independent of one another.

4. The causal forces so operate that deviation above the population mean and below it are balanced as to the magnitude and number. This is the condition of symmetry.

4.4.5 Constants of Normal Distribution

In the normal distribution the values of different constants are:

Mean **X** or μ

Standard Deviation a

First Moment $u_1 = 0$

Second Moment $\mu_2 = \sigma^2$

Third Moment $\mu_3 = 0$

Fourth Moment $\mu_4 = 3\sigma^4 = 3\mu^2$

B or Moment coefficient of skewness = $\mu_3^2 / \mu_2^3 = 0$

B or Moment coefficient of kurtosis = $\mu_4/\mu_3^2 = 3\mu_2^3/\mu_2^3 = 3\sigma^4/\sigma^4 = 3$

4.4.6 Finding Area under the Normal Curve

The equation of the normal curve gives the ordinate of the curve corresponding to any given value of x. Though it is important to define the normal distribution in the form of an equation in order to observe the relationship between x values and corresponding y values, yet in most applications in statistical inference we are not interested in the ordinate; of the curve. It has been pointed out earlier that the normal distribution is continuous, and in continuous distribution, the random variable can assume an infinite number of values in an interval. In order to computer the probability of a random variable lying between two specified values, we need to know the area under the normal curve between two parameters, the mean (□) and standard deviation (□). Since u and a can assume an infinite number of values, it is impossible to tabulate the areas under the curve for different values of □ and □. A normal distribution or □ = 0 and standard deviation = 1 is called the standard normal distribution or the unit normal distribution, as given, by the equation. Such a normal curve with zero mean and unit standard deviation is called as the standard normal curve.

For convenience, it is useful to transform a normally distributed variable into such a form that a single table of areas under the normal curve would be applicable regardless of the units of the original data. We need to know the area between the mean and n point above the mean some specified distance measured in standard deviation. Because the distance will vary with the situation, it is treated as a variable and denoted by the letter z sometimes the value of z is referred to as a normal deviate. The distance z that separates a possible normal random variable value x from its mean may be determined from the following expression for the normal deviate:

$$z = [x - \mu]/\sigma$$

Wherez = z transformation

x = thevalue of the observation

μ = the mean of the distribution

σ = the standard of deviation of the distribution.

Suppose, for a distribution the mean is 300 and standard deviation is 20. The value of z, when x = 300 will be: x -μ /σ or 300 -300/20 = 1 and x value 340 is equivalent to a value of 2, since 340 - 300/2=2 All other z-scale values are obtained in a similar manner.

The following examples will show how the Table of Area under Normal Curve is consulted to find the area under the normal curve.

Examples:

(a) Find the area under the normal curve between Z = 0 and Z = 1.54.

(b) Find the area under the normal curve between Z = –1.5 and Z = 0.

(c) Find the area between Z = –0.45 and Z = 2.5

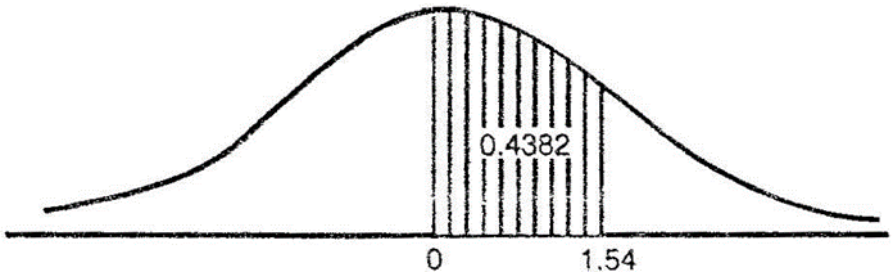
(d) Find the area to the right of Z = +0.36

(e) Find the area to the right of Z = – 1.25 or greater than Z = –1.25.

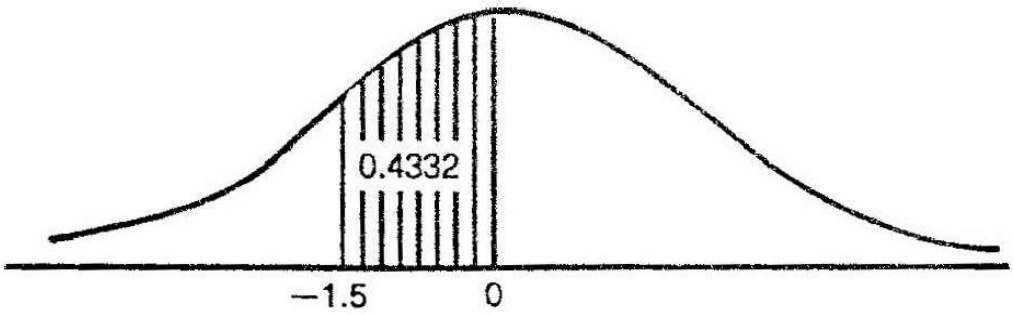
(f) Find the area to the right of Z = +1 and to the left of Z = –1.

Solutions:

(a) The entry corresponding to z = 1.54 is 0.4382 and this gives the shaded area in the following figure between z = 0 and z = 1.54.



(b) The table given at the end of the book does not contain entries corresponding to negative values of

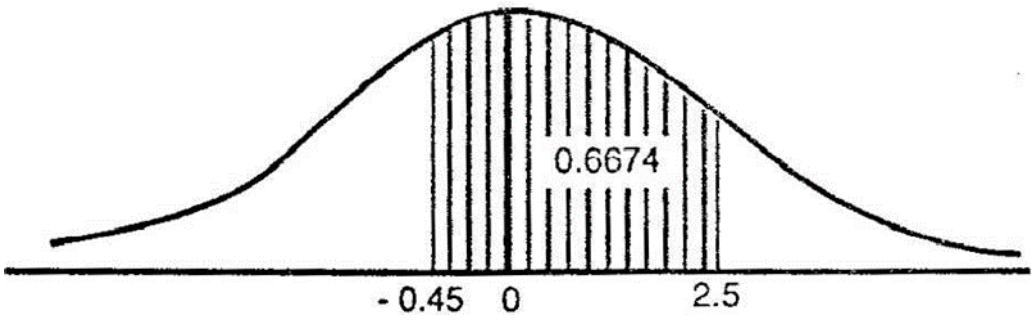


Z. But since the curve is symmetrical, we can find the area between Z = 0 and Z = – 1.5 by looking the area corresponding to Z = 0 and Z = –1.5. Therefore, the entry corresponding to Z = 1.5 is 0.4332 and it measures the shaded area in the following figure between Z = 0 and Z = – 1.5.

(c) Required Area = (Area between Z = –0.45 and Z = 0) + (Area between Z = 0 and Z = 2.5)

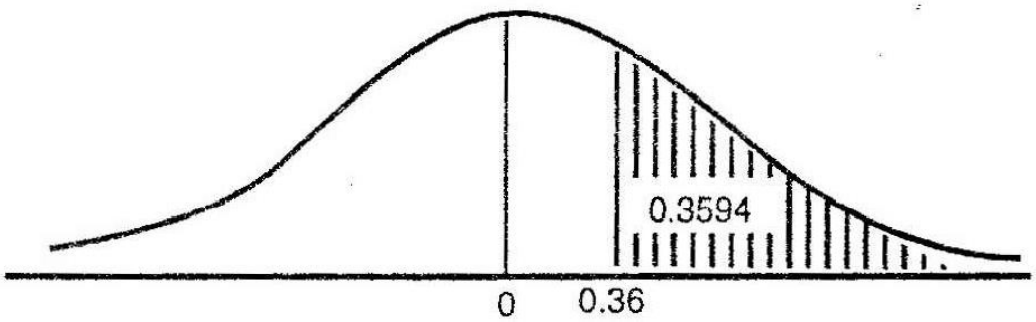
= (Area between Z = 0 and Z = 0.45) + (Area between Z = 0 and Z = 2.5)

= 0.1736 + 0.4938 = 0.6674.



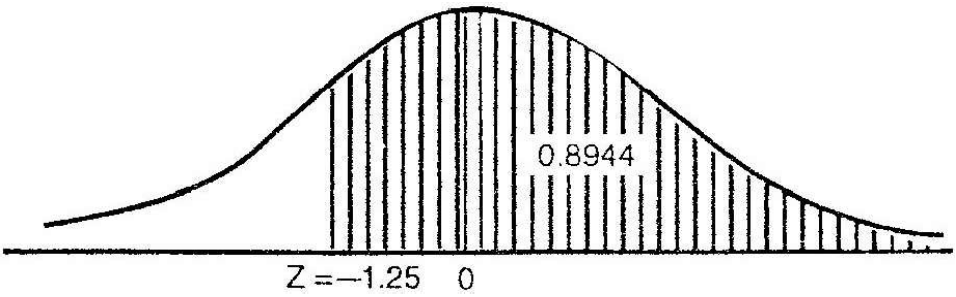
(d) Required Area = (Area to the right of Z = 0) – (Area between Z= 0 and Z = 0.36)

= .5000 – .1406 = 0.3594



(e) Required Area = (Area between Z= – 1.25 and Z= 0)+(Area to the right of Z= 0)

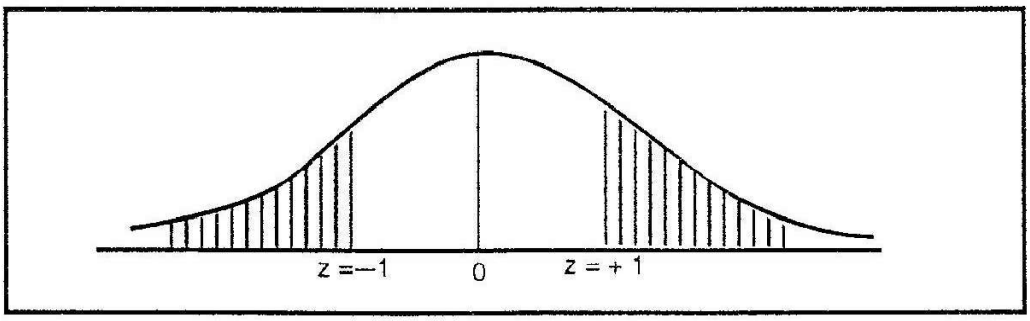
= 0.3944 + .5000



= 0.8944

(f) Required Area= Total Area–(Area between Z = –1 and Z=0)–(area between Z =0 and Z = +1)

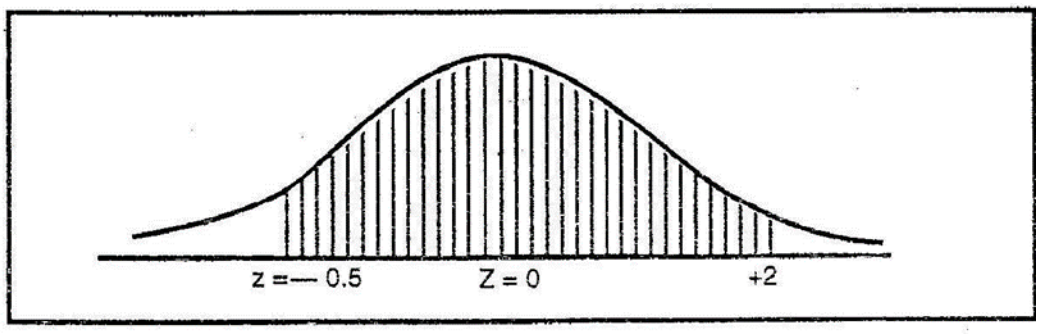
= 1 – 0.3413 – .3413



= 1 – 0.6826 = 0.3174

Illustration 4.8: A normal curve has $\bar{X}$ = 20 and □ = 10. Find the area between X_1 = 15 and X_2 = 40.

Solution: Given $\bar{X}$ = 20, □ = 10



$$Z_1 = \text{SNV corresponding to } 15 = \frac{X_1 - \bar{X}}{\sigma} = \frac{15 - 20}{10} = \frac{-5}{10} = -0.5$$

$$Z_2 = \text{SNV corresponding to } 40 = \frac{X_2 - \bar{X}}{\sigma} = \frac{40 - 20}{10} = \frac{20}{10} = +2.0$$

Required area = Area between Z = –.5 and Z = 0
+ Area between Z = 0 and Z = +2
= 0.1915 + 0.4772 = 0.6687.

Example 4.9: An aptitude test for selecting officers in a bank was conducted on 1,000 candidates, the average score is 42 and the standard deviation of scores is 24.

Assume normal distribution for the scores, find

(i) the number of candidates whose score exceed 60

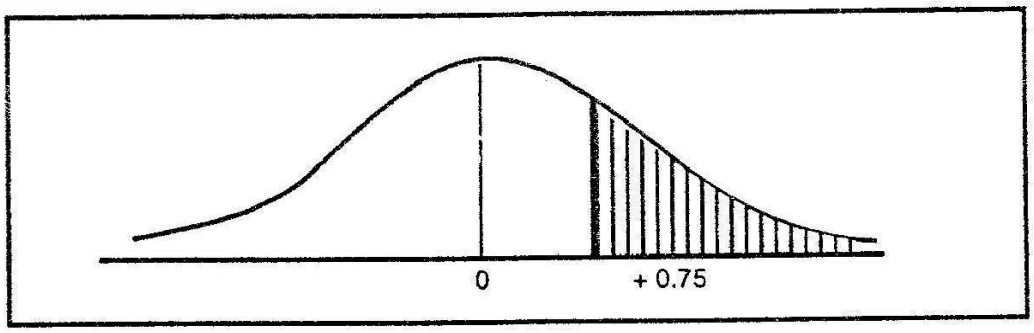
(ii) the number of candidates whose score lie between 30 and 66.

Solution: Given $\overline{X} = 42, \sigma = 24, N = 1000$

(i) Exceeding 60

$$Z = \text{SNV corresponding to } 60 = \frac{X - \overline{X}}{\sigma} = \frac{60 - 42}{24}$$

$$= \frac{18}{24} = \frac{3}{4} = +0.75$$



Required Proportion = Area to the right of Z = 0 – Area between Z = 0 and Z = +0.75

= 0.5000 – 0.2734 = 0.2266

Number of candidates whose score exceeds 60 = 1,000 × 0.2266

= 226.6 □ 227

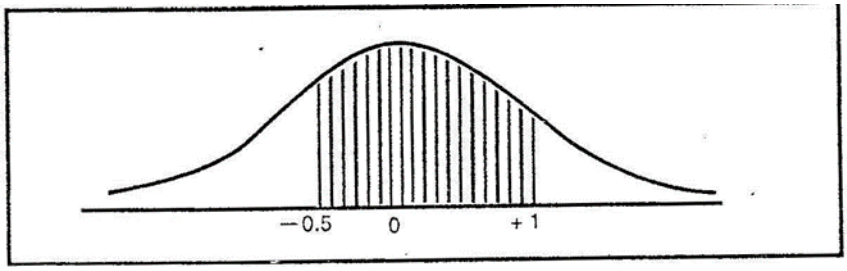
(ii) Between 30 and 66

$$Z_1 = \text{SNV corresponding to } 30 = \frac{X_1 - \overline{X}}{\sigma} = \frac{30 - 42}{24}$$

$$= \frac{-12}{24} = -0.5$$

$$Z_2 = \text{SNV corresponding to } 66 = \frac{X_2 - \overline{X}}{\sigma} = \frac{66 - 42}{24}$$

$$= \frac{24}{24} = +1$$



Required Proportion = Area between Z = –0.5 and Z = 0 +

Area between Z = 0 and Z = +1 = 0.1915 + 0.3413 = 0.5328

Number of candidates whose score lie between 30 and 66

= 1,000 × 0.5328 = 532.8 or 533

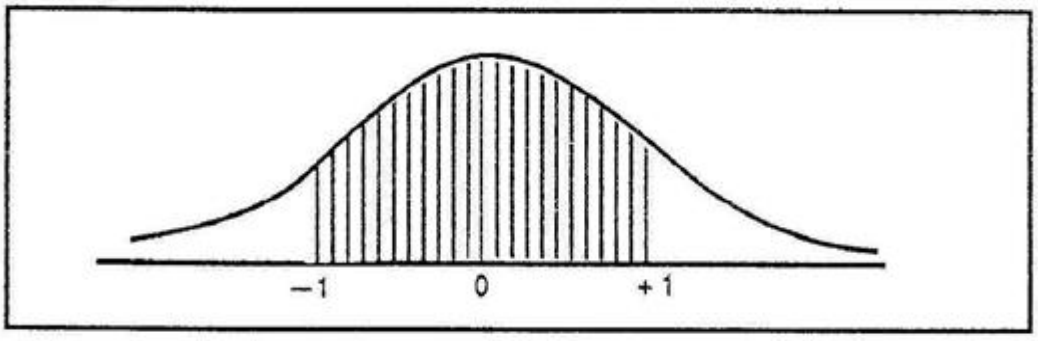
Illustration 4.10: In an entrance test for admission 900 students appeared. Their average marks were 50 and standard deviation 20. Asuming normal distribution find (i) the number of students securing between 30 and 70 (ii) the number of students exceeding the score of 65.

Solution: Given $\bar{X}$ = 50, σ = 20, ℕ = 900.

(i) Between 30 and 70

$$\text{For } X_1 = 30,\ Z_1 = \frac{X_1 - \bar{X}}{\sigma} = \frac{30 - 50}{20} = \frac{-20}{20} = -1$$

$$\text{For } X_2 = 70,\ Z_2 = \frac{X_2 - \bar{X}}{\sigma} = \frac{70 - 50}{20} = \frac{20}{20} = +1$$



Required Proportion = Area between Z = –1 and Z = 0

+ Area between Z = 0 and Z = + 1

= 0.3413 + 0.3413

= 0.6826

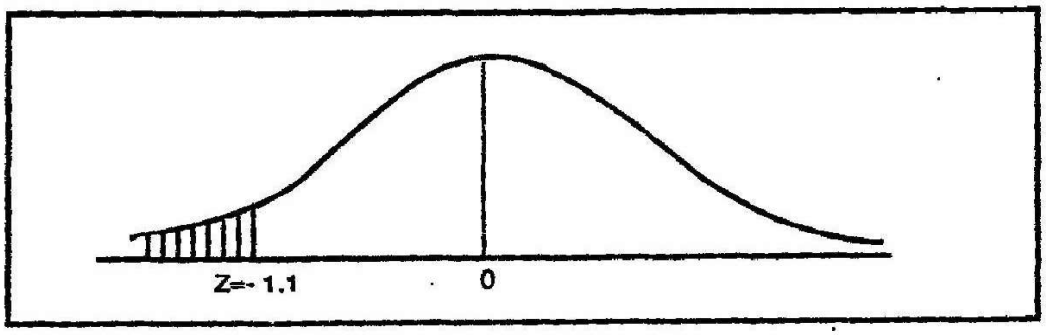
Number of students securing between 30 and 70 = 0.6826 × 900 = 614.34 = 614

Illustration 4.11: Net profit of 400 companies is normally distributed with a mean profit of Rs. 150 lakhs and a standard deviation of Rs. 20 lakhs. Find the number of companies whose profits (Rs. lakhs) are (i) less than 128, (ii) more than 175.

Solution: Given N = 200, $\bar{X}$ = 150, σ = 20

(i) Less than 128

For X = 128, Z = $\frac{128-150}{20} = \frac{-22}{20}$ = -1.1



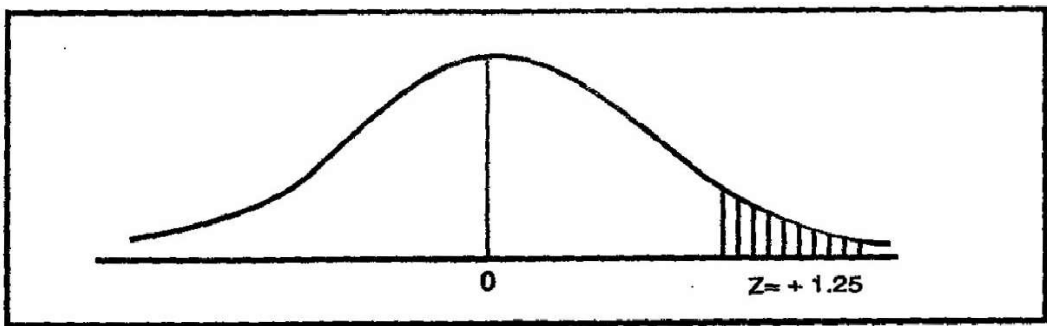
Required Proportion = Area to the left of Z = 0

– Area between Z = – 1.1 and Z = 0

= .5000 – .3643 = 0.1357

(i) More than 175

For X = 175, Z = $\frac{175 - 150}{20} = \frac{25}{20}$ = 1.25



Required Proportion = Area to the right of Z = 0 – Area between Z = 0 and Z = 1.25

= .5000 – .3944 = 0.1056

No. of companies = .1056 × 400 = 42.24 = 42 app.

Illustration 4.12

(a) The mean length of steel bars produced by a company is 10 meters and the standard deviation 20 cms. 5000 bars are purchased by a building contractor. How many of these bars are expected to be shorter than 9.75 meters in length?

(b) For a set of 1000 observations known to be normally distributed the mean is 534 cms. and S.,D. is 13.5 cms. How many observations are likely to exceed 561 cms.? How many will be between 520.5 cms. and 547.5 cms.?

Solution:

(a) $z = x - \mu/\sigma$, 9.75 - 10/.20 = -1.25

z = - 1.25 covers the area .3944

.5000 - .3944 = .1056 x 5000 = 528

528 bars are expected to be shorter than 9.75 meters.

(b) $z = x - \mu/a$, $561 - 534/13.5 = 2$

$z = 2$ covers the area + .4772

The area above 561 will be .5000 - .4772 = =.0228

.0228 x 1000 = 22.8 or 23 observations are likely to exceed 561 cms.

$z = x - \mu/\sigma$, $547.5 - 534/13.5 = 1$

$520.5 - 534/13.5 = - 1$

The area between 0 to 1 = .3413

The area between 0 to 1 = .3413

Total area = .6826 x 1000 = 682.6 or 683 observations lie between 520.5 and 547.5.

Illustration 4.13

Fit a normal distribution to the following data:

Class:	10-20	20-30	30-40	40-50	50-60	60-70	70-80	Total
Frequency:	2	11	24	33	20	8	2	100

Solution:

Since the mean and standard deviation of the data are not given, first these should be calculated.

Class	Mid-values m	Frequency f	d = M - 45/10	fd	fd^2
10-20	15	2	- 3	- 6	18
20-30	25	11	- 2	- 22	44
30-40	35	24	- 1	- 24	24
40-50	45	33	0	0	0

50-60	55	20	+ 1	+ 20	20
60-70	65	8	+ 2	+ 16	32
70-80	75	2	+ 3	+ 6	18
		N = 100		- 10	156

$X = A + fd/N \times I = 45 - 10/100 \times 10 = 44$

$\sigma = df^2/N - (fd/N)^2 \times 1 = 156/100 - (- 0/100)^2 \times 10$

$= 1.56 - .01 \times 10$

$= 1.55 \times 10 = 12.5$

The Mean ordinate = $0.3989 \times Ni/\sigma$

$= .3989 \times 100 \times 10/12.5 = 31.912$

Class	Mid-values m	(m - X) x	x/σ	Proportionate ordinate	Height of ordinate	
				$e - \sigma/2(x/\sigma)^2$	or	
					Expected	
					Frequency	
1	2	3	4	5	6	
10-20	.15	- 29	- 2.39	.05750	1.84 or	2
20-30	25	- 19	- 1.52	.31500	10.05 or	10
30-40	35	- 9	- 0.72	.77167	21.63 or	24
40-50	45	+ 1	+ 0.08	.99685	31.81 or	32
50-60	55	+ 11	+ 0.88	.67896	21.67 or	22

60-70	65	+ 21	+ 1.68	.24385	7.78 or	8
70-80	75	+ 31	+ 2.48	.04618	1.47 or	2
					100	

Steps involved in computing Heights of Ordinates

The following steps are involved in computing the heights of ordinates:

(1) The mid-points of the various classes are obtained. (column 2)

(2) Deviations are taken of each mid-point from the mean (m -X), and they are denoted by x. (column 3).

(3) Each value of x is divided by the a, (x/σ). (column 4)

(4) Height of distance corresponding to each value of x/a are obtained from the table ordinates of the Normal Curve. (column 5)

(5) Each figure of column 5 is multiplied by the value of mean ordinate. The resultant figures are the values of the heights of ordinate at various distances from the mean, i.e., the expected frequencies of the normal distribution.

4.5 CHECK YOUR PROGRESS

1. The mean of the binomial distribution is ----------.
2. The variance of the binomial distribution is always ----------.
3. Shape of the normal curve can be related to ----------.
4. Total Area under the normal curve is ----------.
5. The F-distribution curve in respect of tails is ----------.

4.6 SUMMARY

A random variable can be either discrete or continuous. A random variable is said to be discrete if the set of values defined by it over the sample space is finite. On the other hand, a random variable is 'continuous' if it can assume any (real) value in an interval. If the random variable X is a discrete one,

the probability function P(X) is called 'probability mass function' and its distribution as 'discrete probability distribution' and if the random variable X is of continuous type, then the probability function P(X) is called 'probability density function' and its distribution as 'continuous probability distribution.'

Knowledge of the expected behaviour of a phenomenon or, in other words, the expected frequency distribution is of great help in a large number of problem in practical life. They serve as benchmarks against which we compare observed distributions and act as substitutes for actual distributions when the latter are costly to obtain or cannot be obtained at all. They provide decision-makers with a logical basis for making decisions and are useful in making predictions on the basis of limited information or probability considerations. For example the proprietor of a soe store must know something about the distribution of the size of his potential customers' feet; otherwise, he may find himself with huge stock of shoes which have no market. Similarly, the manufacture of ready made garments must know the sizes of collars for which he expects maximum demand school, college or university should know what they expect of the students. It is only then that they would be in a position to comment on good or bad performance.

Amongst probability or expected frequency distributions, the following three are more popular:

1. Binomial Distribution.
2. Poisson Distribution.
3. Normal Distribution.

Among these the first two distributions are of discrete type and the last one of continuous type. It may also be pointed out that of the three distributions mentioned; the Binomial, Poisson and Normal find much wider applications in practice.

4.7 KEYWORDS

Binomial distribution: A discrete probability distribution of outcomes of an experiment known as a Bernoulli process.

Continuous random variable: A variable that is allowed to take on any value within a given range.

Discrete random variable: A variable that is allowed to take on only integer values.

Continuous probability distribution: A probability distribution in which the random variable is permitted to take on any value within a given range.

Discrete probability distribution: A probability distribution in which the random variable is permitted to take on only integer values.

Expected value of a random variable: A weighted average obtained by multiplying each possible value of the random variable with its probability of occurrence as a weight.

Normal distribution: A continuous probability distribution in which the curve is bell- shaped having a single peak. The mean of the distribution lies at the center of the curve and the curve is symmetrical around a vertical line erected at the mean. The tails of the curve extend indefinitely parallel to the horizontal axis.

Poisson distribution: A discrete probability distribution in which the probability of occurrence of an event within a very small time period is very small, and the probability that two or more such events will occur within the same small time interval is negligible. The occurrence of an event within one time period is independent of the other.

Standard normal probability distribution: A normal probability distribution with mean mean equal to zero and standard deviation equal to one.

4.8 SELF ASSESSMENT QUESTIONS

1. What is meant by probability frequency distribution? Discuss the salient features of the Binomial, Normal and Poisson distributions.

2. (a) Explain what is meant by binomial distribution and obtain the mean and standard deviation of such a distribution.

   (b) State the condition under which binomial probability model is appropriate.

3. (a) Write a critical note on the role of normal distribution in statistics.

   (b) Define the binomial distribution N(p+q)n and show that its mean is np and standard deviation npq.

4. When does Binomial distribution tend to become (i) a normal, (ii) Poisson distribution? Explain clearly.

5. (a) Give the expressions for the mean and the standard deviation of the binomial and Poisson distribution, explaining the meaning of each symbol you use.

   (b) Explain the general characteristics of a Poisson distribution. Give three examples familiar to you, the distribution of which will conform to the Poisson form.

6. (a) Why does the normal distribution hold the most honourable position in probability theory?

   (b) Differentiate between Binomial. Poisson and Normal distribution. Of which type would you expect the distribution to be in the following cases and why?

   (i) Frequencies of experiments with dice throwing and

   (ii) Frequencies of suicides in the total population of India, suicide being a rare phenomenon.

7. (a) How does a normal distribution differ from a binomial distribution? What are the important properties of a normal distribution? How are they useful in random sampling investigation?

   (b) Discuss the various types of relationships which are found in a normal distribution. What is meant by area relationship in a normal curve and what use it is in the theory of sampling?

8. (a) Find the mean and variance of Poisson distribution.

   (b) How are the value of mean and standard deviation calculated in Binomial and Poisson distribution?

9. (a) Calculate the ordinates of the binomial: $128\ (1/2 + 1/2)^4$.

   (b) Write a critical note on the role of normal distribution in statistics.

   (c) Determine the binomial distribution, in which $B_1 = 1/36$ and $B_2 = 35/12$.

$$^{18}C\ (1/3)^r{}_r(2/3)^{18 - r},\ r = 0, 1, 2\ldots..18]$$

10. In any army battalion 2/5 of soldiers are known to be married and the remainder 2/5 unmarried. Calculate the probability of getting 0, 1, 2….5 married soldiers in a row of 5 soldiers. If 500 rows

each of 5 soldiers are standing on a ground, approximately how many rows are expected to contain (i) all married soldiers, and (ii) all unmarried soldiers?

[32/3125, 240/3125, 720/3125, 1000/3125, 243/3125. (i) 39, (ii) nearly 5]

11. Five dice were thrown together 96 times. The number of times 4, 5 or 6 was actually thrown in the experiment is given below. Calculates the expected frequencies.

No. of dice showing 4, 5 or 6:	0	1	2	3	4	5
Observed Frequency:	1	10	24	35	18	8

[Expected frequency according to Binomial Distribution : 3, 15, 30, 30, 15,3]

4.9 ANSWERS TO CHECK YOUR PROGRESS

1. np
2. Less than mean
3. Bell
4. One
5. Positively Skewed

4.10 REFERENCES/SUGGESTED READINGS

1. Hooda, R P: Statistics for Business and Economics, 3rd Edition, MacMilan India Ltd., New Delhi.
2. Gupta, S P: Statistical Methods, 7th Edition, Sultan Chand and Sons, New Delhi. Bhardwaj, R S: Business Statistics, Excel Book, New Delhi.
3. Murray R. Spiegel and Larry J. Stephens, Statistics, 3rd Edition, TMH, New Delhi.
4. Viswanathan, PK, Business Statistics, First edition, Pearson Education (Singapore) Ltd., Delhi.

NOTES

NOTES

NOTES

NOTES