8. REGRESSION ANALYSIS

NOTES

STRUCTURE

8.1. INTRODUCTION

In the discussion of correlation, we estimated the degree of relationship between variables. The coefficient of correlation r, $(-1 \le r \le 1)$ measured the *degree* of relationship between variables. A numerically high value of 'r' resulted because of closeness of relation between the variables, under consideration. The coefficient of correlation is unable to depict the *nature* of relationship between the variable. For a given data regarding the corresponding values of two related variables, the coefficient of correlation cannot give the *estimated value* of a variable, corresponding to a certain value of the other related variable. For example, the coefficient of correlation between 'height' and 'weight' of a group of students of a university cannot help to give the estimated weight (resp. height) of a student with given height (resp. weight). This type of assignment is dealt with the tools of *regression analysis*.

8.2. MEANING

The literal meaning of the word 'regression' is 'stepping back towards the average'. British biometrician *Sir Francis Galton* (1822–1911) studies the heights of many persons and concluded that the offspring of abnormally tall or short parents tend to *regress* to the average population height. In statistics, *regression analysis* is concerned with the measure of average relationship between variables. Here we shall deal with

NOTES

the derivation of appropriate functional relationships between variables. Regression explains the nature of relationship between variables.

There are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable (or regressed variable or predicted variable or explained variable).* The variable which influences the value of dependent variable is called *independent variable (or regressor or predictor or explanator).* Prediction is possible in regression analysis, because here we study the average relationship between related variables.

8.3. USES OF REGRESSION ANALYSIS

The tools of regression analysis are definitely more important and useful than those of correlation analysis. Some of the important uses of regression analysis are as follows:

(*i*) Regression analysis helps in establishing relationship between dependent variable and independent variables. The independent variables may be more than one. Such relationships are very useful in further studies of the variables, under consideration.

(*ii*) Regression analysis is very useful for prediction. Once a relation is established between dependent variable and independent variables, the value of dependent variable can be predicted for given values of the independent variables. This is very useful for predicting sale, profit, investment, income, population, etc.

(*iii*) Regression analysis is specially used in Economics for estimating demand function, production function, consumption function, supply function, etc. A very important branch of Economics, called *Econometrics,* is based on the techniques of regression analysis.

(*iv*) The coefficient of correlation between two variables can be found easily by using the regression lines between the variables.

8.4. TYPES OF REGRESSION

If there are only two variables under consideration, then the regression is called **simple regression.** For example, the study of regression between 'income' and 'expenditure' for a group of family would be termed as simple regression. If there are more than two variables under consideration then the regression is called **multiple regression.** In this text, we shall restrict ourselves to the study of only simple regression. The regression is called **partial regression** if there are more than two variables under consideration and relation between only two variables is established after excluding the effect of other variables. The simple regression is called **linear regression** if the point on the scatter diagram of variables lies almost along a line otherwise it is termed as **non-linear regression** or **curvilinear regression.**

8.5. REGRESSION LINES

Let the variables under consideration be denoted by 'x' and 'y'. The line used to estimate the value of y for a given value of x is called the *regression line* of *y on x*. Similarly, the

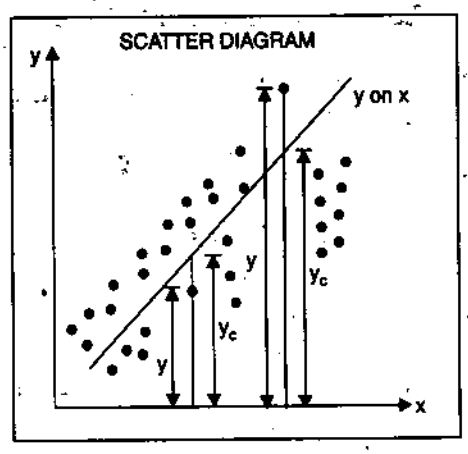
NOTES

line used to estimate the value of x for a given value of y is called the *regression line of x on y*. In regression line of y on x (x on y), the variable y is considered as the dependent (independent) variable whereas x is considered as the independent (dependent) variable. The position of regression lines depends upon the given pairs of value of the variables, Regression lines are also known as *estimating lines*. We shall see that in case of perfect correlation between the variables, the regression lines will be coincident. The angle between the regression lines will increases for 0° to 90° as the correlation coefficient numerically decreases from 1 to 0. If for a particular pair of variables, $r = 0$, then the regression lines will be perpendicular to each other. The regression lines will be determined by using the *principle of least squares*.

8.6. REGRESSION EQUATIONS

We have already noted that for two variables x and y, there can be two regression lines. If the intention is to depict the change in y for a given change in x, then the regression line of y on x is to be used. Similar argument also works for regression line of x on y.

(*i*) **Regression equation of y on x.** The regression equation of y on x is estimated by using the 'principle of least squares'. This principle will ensure that the sum of the squares of the *vertical* deviations of actual values of y from estimated values for all possible values of x is minimum.



Mathematically, $\Sigma(y - y_c)^2$ is least, where y and y_c are the corresponding actual and computed values of y for a particular value of x.

Let n pairs of values $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ of two variables x and y be given.

Let the regression equation of y on x be $y = a + bx$. ...(1)

By using derivatives, it can be proved that the constants a and b are found by using the *normal equations:*

$$\Sigma y = an + b\Sigma x \quad ...(2)$$

and $$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad ...(3)$$

Dividing (2) by n, we get

$$\frac{\Sigma y}{n} = a + b\frac{\Sigma x}{n}$$

$$\Rightarrow \quad \bar{y} = a + b\bar{x} \quad ...(4)$$

NOTES

Subtracting (4) from (1), we get

$$y - \bar{y} = b(x - \bar{x}) \quad \ldots(5)$$

Multiplying (2) by Σx and (3) by n and subtracting, we get

$$(\Sigma x)(\Sigma y) - n\Sigma xy = b(\Sigma x)^2 - bn\Sigma x^2$$

$$\Rightarrow \quad n\Sigma xy - (\Sigma x)(\Sigma y) = b(n\Sigma x^2 - (\Sigma x)^2)$$

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}.$$

The constant b is denoted by b_{yx} and is called **regression coefficient** of y on x.

$$\therefore \quad (5) \Rightarrow \quad y - \bar{y} = b_{yx}(x - \bar{x}), \text{ where } b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

Remark. $b_{yx} = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$ implies

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

$$\therefore \quad \mathbf{b_{yx} = r\frac{\sigma_y}{\sigma_x}}.$$

Thus we see that the regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$, where $\bar{x} = \dfrac{\Sigma x}{n}$, $\bar{y} = \dfrac{\Sigma y}{n}$, $b_{yx} = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$, which is also equal to $r\dfrac{\sigma_y}{\sigma_x}$.

Example 8.1. *Find the regression equation of y on x when we know :*

$$\bar{x} = 68.2,\ \bar{y} = 9.9,\ \frac{\sigma_y}{\sigma_x} = 0.44,\ r = 0.76.$$

Solution. We have $\bar{x} = 68.2,\ \bar{y} = 9.9,\ \dfrac{\sigma_y}{\sigma_x} = 0.44,\ r = 0.76.$

The regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$$\Rightarrow \quad y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \Rightarrow \quad y - 9.9 = (0.76)(0.44)(x - 68.2)$$

$$\Rightarrow \quad y - 9.9 = 0.3344\,(x - 68.2) \quad \Rightarrow \quad y = 0.3344x + 9.9 - (0.3344)(68.2)$$

$$\Rightarrow \quad \mathbf{y = 0.3344x - 12.9061.}$$

Example 8.2. *x and y are correlated variables. Ten observations of values of (x, y) have the following results:*

$$\Sigma x = 55,\ \Sigma y = 55,\ \Sigma xy = 350,\ \Sigma x^2 = 385.$$

Predict the value of y when the value of x is 6.

Solution. To predict the value of y for a given value of x, we shall require the equation of regression line of y on x.

The equation of regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x}). \quad \ldots(1)$$

NOTES

We have $\Sigma x = 55, \Sigma y = 55, \Sigma xy = 350, \Sigma x^2 = 385, n = 10$

Now $\bar{x} = \frac{\Sigma x}{n} = \frac{55}{10} = 5.5, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{55}{10} = 5.5$

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{10(350) - (55)(55)}{10(385) - (55)^2} = \frac{475}{825} = 0.5758.$$

$\therefore$ (1) $\Rightarrow$ $y - 5.5 = 0.5758\,(x - 5.5)$

$\Rightarrow$ $y = 0.5758x + 2.3331.$

This is the equation of regression line of y on x.

When $x = 6$, the predicted value of y

$= 0.5758\,(6) + 2.3331 = \mathbf{5.7879.}$

Example 8.3. *For the following data, find the regression line of y on x :*

x	1	2	3	4	5	8	10
y	9	8	10	12	14	16	15

Solution. **Regression line of y on x**

S. No.	*x*	*y*	*xy*	x^2
1	1	9	9	1
2	2	8	16	4
3	3	10	30	9
4	4	12	48	16
5	5	14	70	25
6	8	16	128	64
7	10	15	150	100
$n = 7$	$\Sigma x = 33$	$\Sigma y = 84$	$\Sigma xy = 451$	$\Sigma x^2 = 219$

The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{33}{7} = 4.714, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{84}{7} = 12$$

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(451) - (33)(84)}{7(219) - (33)^2} = \frac{385}{444} = \mathbf{0.867.}$$

$\therefore$ The equation of regression line of y on x is

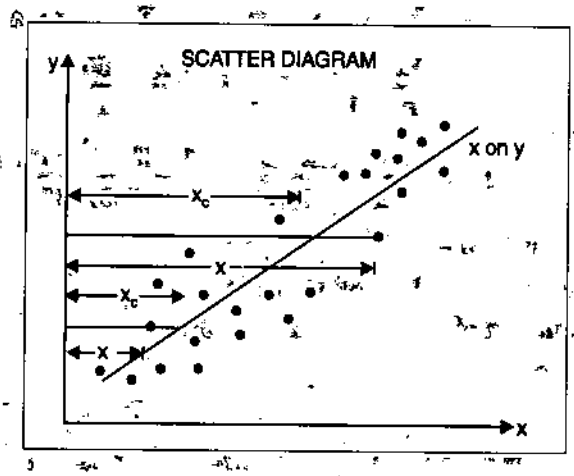
$$y - 12 = 0.867\,(x - 4.714)$$

or $$y = 0.867x + 12 - (0.867)(4.714)$$

or $$\mathbf{y = 0.867x - 7.913.}$$

(*ii*) **Regression equation of x on y.** The regression equation of x on y is also estimated by using the 'principle of least squares'. This principle will ensure that the sum of the squares of the *horizontal* deviations of actual values of x from estimated values for all possible values of y is minimum. Mathematically, $\Sigma(x - \bar{x}_c)^2$ is least, where x and x_c are the corresponding actual and computed values of x for a particular value of y.

NOTES



Let n pairs of values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ of two variables x and y be given.

Let the regression equation of x on y be $x = a + by$ ...(1)

By using derivatives, it can be proved that the constants a and b are found by using the *normal equations*:

$$\Sigma x = an + b\Sigma y \quad \ldots(2)$$

and $$\Sigma xy = a\Sigma y + b\Sigma y^2. \quad \ldots(3)$$

Dividing (2) by n, we get

$$\frac{\Sigma x}{n} = a + b\frac{\Sigma y}{n}$$

$$\Rightarrow \quad \bar{x} = a + b\bar{y} \quad \ldots(4)$$

Subtracting (4) from (1), we get

$$x - \bar{x} = b(y - \bar{y}) \quad \ldots(5)$$

Multiplying (2) by Σy and (3) by n and subtracting, we get

$$(\Sigma x)(\Sigma y) - n\Sigma xy = b(\Sigma y)^2 - bn\Sigma y^2$$

$$\Rightarrow \quad n\Sigma xy - (\Sigma x)(\Sigma y) = b(n\Sigma y^2 - \Sigma y)^2$$

$$\therefore \quad b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$$

The constant b is denoted by b_{xy} and is called **regression coefficient** of x on y.

(5) $\Rightarrow \quad x - \bar{x} = b_{xy}(y - \bar{y})$, where $b_{xy} = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$

Remark $b_{xy} = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$ implies

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

$$\mathbf{b_{xy} = r\frac{\sigma_y}{\sigma_x}}$$

Thus, we see that the regression equation of x on y is $\mathbf{x - \bar{x} = b_{xy}(y - \bar{y})}$, where $\mathbf{\bar{x} = \frac{\Sigma x}{n}}$, $\mathbf{\bar{y} = \frac{\Sigma y}{n}}$, $\mathbf{b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}}$, which is also equal to $\mathbf{r\frac{\sigma_x}{\sigma_y}}$.

Example 8.4. *Find the regression coefficient b_{xy} between x and y for the following data:*

$$\Sigma x = 30,\ \Sigma y = 42,\ \Sigma xy = 199,\ \Sigma x^2 = 184,\ \Sigma y^2 = 318,\ n = 6.$$

Solution. $$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{6(199) - (30)(42)}{6(318) - (42)^2} = \frac{-66}{144} = -0.4583.$$

NOTES

Example 8.5. *For observations of pairs (x, y) of the variables x and y, the following results are obtained:*

$$\Sigma x = 110,\ \Sigma y = 70,\ \Sigma x^2 = 2500,\ \Sigma y^2 = 2000,\ \Sigma xy = 100,\ n = 20.$$

Find the equation of the regression line of x on y. Estimate the value of x when y = 4.

Solution. We have

$$\Sigma x = 110,\ \Sigma y = 70,\ \Sigma x^2 = 2500,\ \Sigma y^2 = 2000,\ \Sigma xy = 100,\ n = 20.$$

The equation of regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y}). \quad \ldots(1)$$

Now $$\bar{x} = \frac{\Sigma x}{n} = \frac{110}{20} = 5.5,\quad \bar{y} = \frac{\Sigma y}{n} = \frac{70}{20} = 3.5$$

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{20(100) - (110)(70)}{20(2000) - (70)^2} = \frac{-5700}{35100} = -0.1624$$

$\therefore$ (1) $\Rightarrow$ $x - 5.5 = -0.1624\,(y - 3.5)$

$\Rightarrow$ $x = -0.1624y + (0.1624)(3.5) + 5.5$

$\Rightarrow$ $\mathbf{x = -0.1624y + 6.0684.}$

This is the regression equation of x on y.

When $y = 4$, the estimated value of y

$$= -0.1624(4) + 6.0684 = \mathbf{5.4188.}$$

Example 8.6. *Find the equations of the line of regression of y on x and x on y for the data:*

x	*5*	*2*	*1*	*4*	*3*
y	*5*	*8*	*4*	*2*	*10*

Solution. **Regression Equations**

S. No.	*x*	*y*	*xy*	x^2	y^2
1	5	5	25	25	25
2	2	8	16	4	64
3	1	4	4	1	16
4	4	2	8	16	4
5	3	10	30	9	100
$n = 5$	$\Sigma x = 15$	$\Sigma y = 29$	$\Sigma xy = 83$	$\Sigma x^2 = 55$	$\Sigma y^2 = 209$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{15}{5} = 3,\quad \bar{y} = \frac{\Sigma y}{n} = \frac{29}{5} = 5.8$$

The regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{5(83) - (15)(29)}{5(55) - (15)^2} = \frac{-20}{50} = -0.4$$

NOTES

∴ The equation is

$$y - 5.8 = -0.4\,(x - 3)$$

or $$y = -0.4x + (0.4)3 + 5.8$$

or $$\mathbf{y = -0.4x + 7.}$$

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

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{5(83) - (15)(29)}{5(209) - (29)^2} = \frac{-20}{204} = -0.098$$

∴ The equation is

$$x - 3 = -0.098\,(y - 5.8)$$

or $$x = 0.098y + (0.098)(5.8) + 3$$

or $$\mathbf{x = -0.098y + 3.5684.}$$

Example 8.7. *You are given below the following information about advertisement and sales:*

	Advt. Expenditure (x) (in crore rupees)	*Sales (y) (in crore rupees)*
Mean	*20*	*120*
S.D.	*5*	*25*

Coefficient of correlation = 0.8.

(i) Calculate the regression equations.

(ii) Find the likely sales when advertisement expenditure is ₹ 25 crores.

(iii) What should be advertisement budget if the company wants to attain sales target of ₹ 150 crores?

Solution. We have

$$\bar{x} = 20,\ \bar{y} = 120,\ \sigma_x = 5,\ \sigma_y = 25,\ r = 0.8$$

(i) The regression equation of y on x is $y - \bar{y} = b_{yx}\,(x - \bar{x})$.

$$\Rightarrow \quad y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}\,(x - \bar{x}) \quad \Rightarrow \quad y - 120 = (0.8)\,\frac{25}{5}\,(x - 20)$$

$$\Rightarrow \quad y - 120 = 4\,(x - 20) \quad \Rightarrow \quad \mathbf{y = 4x + 40.}$$

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

$$\Rightarrow \quad x - \bar{x} = r\,\frac{\sigma_x}{\sigma_y}\,(y - \bar{y}) \quad \Rightarrow \quad x - 20 = (0.8)\,\frac{5}{25}\,(y - 120)$$

$$\Rightarrow \quad x - 20 = 0.16\,(y - 120) \quad \Rightarrow \quad \mathbf{x = 0.16y + 0.8.}$$

(ii) We are to estimate the value of y for a given value of x.

∴ We use regression equation of y on x which is $y = 4x + 40$.

∴ When $x = 25$, the estimated value of $y = 4(25) + 40 = 140$

∴ Estimated sales = **₹ 140 crores.**

(iii) We are to estimate the value of x for a given value of y.

∴ We use regression equation of x on y which is $x = 0.16y + 0.8$

∴ When $y = 150$, the estimated value of $x = (0.16)(150) + 0.8 = 24.8$

∴ Estimated advt. expenditure = **₹ 24.8 crores.**

Example 8.8. *Given:*

	x-series	*y-series*
Mean	*5*	*4*
S.D.	*1.224*	*1.414*

Sum of products of deviations from means of x and y series = 6

Number of items = 8.

(i) Obtain the regression equations.

(ii) Estimate the value of x when y = 5.

Solution. (*i*) We have $\bar{x} = 5$, $\bar{y} = 4$, $\sigma_x = 1.224$, $\sigma_y = 1.414$, $\Sigma(x - \bar{x})(y - \bar{y}) = 6$ and $n = 8$.

Now $$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2}\sqrt{\Sigma(y - \bar{y})^2}} = \frac{6}{n\sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}\sqrt{\frac{\Sigma(y - \bar{y})^2}{n}}}$$

$$= \frac{6}{8\sigma_x\sigma_y} = \frac{3}{4(1.224)(1.414)} = 0.433.$$

Regression equation of y on x is

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

$$\Rightarrow \quad y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \Rightarrow \quad y - 4 = 0.433 \times \frac{1.414}{1.224}(x - 5)$$

$$\Rightarrow \quad y - 4 = 0.5\,(x - 5) \quad \Rightarrow \quad y = 0.5x + 4 - 2.5$$

$$\Rightarrow \quad \mathbf{y = 0.5x + 1.5.}$$

Regression equation of x on y is

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

$$\Rightarrow \quad x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad \Rightarrow \quad x - 5 = 0.433 \times \frac{1.224}{1.414}(y - 4)$$

$$\Rightarrow \quad x - 5 = 0.375\,(y - 4) \quad \Rightarrow \quad x = 0.375y + 5 - (0.375) \times 4$$

$$\Rightarrow \quad \mathbf{x = 0.375y + 3.5.}$$

(*ii*) When $y = 5$, the estimated value of $x = (0.375)\,5 + 3.5 = \mathbf{5.375.}$

(By using regression equation of x on y)

Example 8.9. *You are given the following data:*

$\Sigma x = 300$, $\bar{x} = 50$, $\Sigma y = 240$, variance of $x = 2.56$, variance of $y = 1.96$, coefficient of correlation between x and $y = + 0.6$.

Find :

(i) Two regression coefficients

(ii) Two regression equations.

Solution. We have

$$\Sigma x = 300, \quad \bar{x} = 50, \Sigma y = 240, \quad \sigma_x^{\,2} = 2.56 \quad \sigma_y^{\,2} = 2.25\,, r = + 0.6.$$

$$\bar{x} = \frac{\Sigma x}{n} \qquad \Rightarrow \quad 50 = \frac{300}{n} \quad \Rightarrow \quad n = \frac{300}{50} = 6$$

NOTES

$\sigma_x^2 = 2.56 \quad \Rightarrow \quad \sigma_x = +\sqrt{2.56} = 1.6$

$\sigma_y^2 = 1.96 \quad \Rightarrow \quad \sigma_y = +\sqrt{1.16} = 1.4.$

(*i*) $b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.6 \times \frac{1.4}{1.6} = 0.525.$

and $b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.6 \times \frac{1.6}{1.4} = 0.686.$

(*ii*) Regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$\Rightarrow \quad y - 40 = 0.525\,(x - 50) \qquad \left(\bar{y} = \frac{\Sigma y}{n} = \frac{240}{6} = 40\right)$

$\Rightarrow \quad y = 0.525x + 40 - (0.525)\,50$

$\Rightarrow \quad \mathbf{y = 0.525x + 13.75.}$

Regression equation of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$.

$\Rightarrow \quad x - 50 = 0.686\,(y - 40)$

$\Rightarrow \quad x = 0.686y + 50 - (0.686)(40)$

$\Rightarrow \quad \mathbf{x = 0.686y + 22.56.}$

Example 8.10. *From the following data, find the regression equations:*

x	6	2	10	4	8
y	9	11	5	8	7

Solution. **Estimation of Regression Equations**

S. No.	x	y	xy	x^2	y^2
1	6	9	54	36	81
2	2	11	22	4	121
3	10	5	50	100	25
4	4	8	32	16	64
5	8	7	56	64	49
$n = 5$	$\Sigma x = 15$	$\Sigma y = 40$	$\Sigma xy = 214$	$\Sigma x^2 = 220$	$\Sigma y^2 = 340$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{30}{5} = 6, \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{40}{5} = 8.$$

The regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{5(214) - (30)(40)}{5(220) - (30)^2} = \frac{-130}{200} = -0.65.$$

$\therefore$ The equation is

$$y - 8 = -0.65\,(x - 6) \quad \text{or} \quad y = -0.65x + (0.65)6 + 8$$

or $\mathbf{y = 11.9 - 0.65x.}$

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

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{5(214) - (30)(40)}{5(340) - (40)^2} = \frac{-130}{100} = -1.3.$$

NOTES

∴ The equation is

$x - 6 = -1.3(y - 8)$ or $x = -1.3y + (1.3)8 + 6$

or $\mathbf{x = 16.4 - 1.3y.}$

Example 8.11. *In order to find the correlation coefficient between two variables X and Y from 12 pairs of observations, the following calculation were made:*

$$\Sigma X = 30,\ \Sigma X^2 = 670,\ \Sigma Y = 5,\ \Sigma Y^2 = 285,\ \Sigma XY = 344.$$

On subsequent verification, it was discovered that the pair (X = 11, Y = 4) was copied wrongly, the correct values being X = 10 and Y = 14. After making necessary correction, find the:

(i) regression coefficients

(ii) regression equations and

(iii) correlation coefficient.

Solution. We have $\Sigma X = 30,\ \Sigma X^2 = 670,\ \Sigma Y = 5,\ \Sigma Y^2 = 285,\ \Sigma XY = 344.$

Incorrect pair = (X = 11, Y = 4)

Correct pair = (X = 10, Y = 14)

Corrected sums

$$\Sigma X = 30 - 11 + 10 = 29$$

$$\Sigma Y = 5 - 4 + 14 = 15$$

$$\Sigma X^2 = 670 - (11)^2 + (10)^2 = 649$$

$$\Sigma Y^2 = 285 - (4)^2 + (14)^2 = 465$$

$$\Sigma XY = 344 - (11 \times 4) + (10 \times 14) = 440.$$

(i) Regression coefficients

$$b_{YX} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{12(440) - (29)(15)}{12(649) - (29)^2} = \frac{4845}{6947} = \mathbf{0.6974}$$

$$b_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma Y^2 - (\Sigma Y)^2} = \frac{12(440) - (29)(15)}{12(465) - (15)^2} = \frac{4845}{5355} = \mathbf{0.9048.}$$

(ii) Regression equations

$$\bar{X} = \frac{\Sigma X}{n} = \frac{29}{12} = 2.42,\quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{15}{12} = 1.25$$

Regression equation of Y on X is $Y - \bar{Y} = b_{YX}(X - \bar{X})$.

$\Rightarrow$ $Y - 1.25 = 0.6974\ (X - 2.42)$

$\Rightarrow$ $Y = 0.6974X + 1.25 - (0.6974)(2.42)$

$\Rightarrow$ $\mathbf{Y = 0.6974X - 0.438.}$

Regression equation of X on Y is $X - \bar{X} = b_{XY}(Y - \bar{Y})$.

$\Rightarrow$ $X - 2.42 = 0.9048\ (Y - 1.25)$

$\Rightarrow$ $X = 0.9048\ Y + 2.42 - (0.9048)(1.25)$

$\Rightarrow$ $\mathbf{X = 0.9048Y + 1.289.}$

(iii) $$b_{YX} \cdot b_{XY} = \left(r\frac{\sigma_Y}{\sigma_X}\right)\left(r\frac{\sigma_X}{\sigma_Y}\right) = r^2$$

$$\therefore \quad r = \pm\sqrt{b_{YX} \cdot b_{XY}}$$

NOTES

$$\therefore \quad r = \pm \sqrt{0.6974 \times 0.9048} = \pm 0.7944$$

$$b_{YX} > 0 \quad \Rightarrow \quad r > 0. \qquad \left(\because \frac{\sigma_y}{\sigma_x} > 0\right)$$

$$\therefore \quad r = +0.7944.$$

EXERCISE 8.1

1. Find b_{yx} from the following data:

   $\Sigma x = 30$, $\Sigma y = 42$, $\Sigma xy = 199$, $\Sigma x^2 = 184$, $\Sigma y^2 = 318$, $n = 6$.

2. Find b_{yx} from the following data:

x	1	2	3	4	5
y	6	8	7	6	8

3. The following results were worked out from scores in Statistics and Mathematics in a certain examination:

	Scores in Statistics (x)	*Scores in Mathematics (y)*
A.M.	39.5	47.5
S.D.	10.8	17.8

   Karl Pearson's coefficient of correlation is 0.42. Find both regression lines. Estimate the value of y when $x = 50$ and x when $y = 30$.

4. From the following data find the yield of wheat in kg per unit area when the rain fall is 9 inches:

	Mean	*S.D.*
Yield of wheat per unit area (in kg)	10	8
Annual rainfall (in inches)	8	2

   Coefficient of correlation = 0.5.

5. You are given below the following information about advertisement expenditure and sales:

	Advt. Expenditure (x) (in crore rupees)	*Sales (y) (in crore rupees)*
Mean	10	90
S.D.	3	12

   Coefficient of correlation = + 0.8.

   (*i*) Calculate two regression equations.

   (*ii*) Find the likely sales when advertisement expenditure is ₹ 30 crores.

   (*iii*) What should be the advertisement budget if the company wants to attain sales target of ₹ 150 crores?

6. Find the equations of regression lines for the following pairs (x, y) for variables x and y:

   (1, 2), (2, 5), (3, 3), (4, 8), (5, 7).

7. For the following data, determine the regression lines:

x	6	2	10	4	8
y	9	11	5	8	7

From these regression lines, estimate the value of:

(*i*) y when $x = 5$ and (*ii*) x when $y = 10$.

8. Find the regression lines for the following pairs (x, y) for variables x and y:

(1, 6), (5, 1), (3, 0), (2, 0), (1, 1), (1, 2), (7, 1), (3, 5).

9. By using the following data regarding pairs (x, y) for variables x and y, find the most likely value of y, when $x = 6.2$:

(1, 9), (2, 8), (3, 10), (4, 12), (5, 11), (6, 13), (7, 14), (8, 16), (9, 15).

10. A computer while calculating the correlation coefficient between two variables x and y obtained the following constants:

$n = 25$, $\Sigma x = 127$, $\Sigma y = 100$, $\Sigma x^2 = 650$, $\Sigma y^2 = 450$, $\Sigma xy = 516$.

It was however, later discovered at the time of checking that it copied down two pairs of observations as:

x	y
8	12
6	8

while the correct values were:

x	y
8	10
6	10

. After making the necessary corrections, find the:

(*i*) regression coefficients (*ii*) regression equations and

(*iii*) correlation coefficient.

Answers

1. – 0.3235 2. 0.2
3. $y = 0.6922x + 20.1581$, $x = 0.2548y + 27.397$, 54.77, 35.04 4. 12 kg
5. (*i*) $x = 0.2y - 8$, $y = 3.2x + 58$, (*ii*) ₹ 154 crores (*iii*) ₹ 22 crores
6. Regression line of y on x : $y = 1.1 + 1.3x$;
   Regression line of x on y : $x = 0.5 + 0.5y$
7. $y = 11.9 - 0.65x$, $x = 16.4 - 1.3y$ (*i*) 8.65 (*ii*) 3.4
8. $y = 2.8745 - 0.3042x$, $x = 3.4306 - 0.2778y$ 9. 13.14
10. (*i*) $b_{yx} = 0.826$, $b_{xy} = 0.095$
    (*ii*) Regression equation of y on x: $y = 0.826x - 1.96$
    Regression equation of x on y: $x = 0.095y + 4.7$
    (*iii*) $r = 0.28$

8.7. STEP DEVIATION METHOD

When the values of x and y are numerically high, the step deviation method is used.

Deviations of values of variables x and y are calculated from some chosen arbitrary numbers, called A and B. Let h be a positive common factor of all deviations $(x - A)$ of items in the x-series. Similarly let k be a positive factor of all deviations $(y - B)$ of items in the y-series. The step deviations are:

$$u = \frac{x - A}{h}, \quad v = \frac{x - B}{k}$$

In practical problems, if we do not bother to divide the deviations by common factors, then these deviations would be thought of as step deviations of items of given series with '1' as the common factor for both series:

NOTES

The equation of regression line of y on x in terms of step deviations is

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

where $$\bar{x} = A + \left(\frac{\Sigma u}{n}\right)h, \quad \bar{y} = B + \left(\frac{\Sigma v}{n}\right)k$$

and $$b_{yx} = b_{vu} \cdot \frac{k}{h} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \cdot \frac{k}{h}$$

The equation of regression line of x on y in terms of step deviations is

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

where $$\bar{x} = A + \left(\frac{\Sigma u}{n}\right)h, \quad \bar{y} = B + \left(\frac{\Sigma v}{n}\right)k$$

and $$b_{xy} = b_{uv} \cdot \frac{h}{k} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} \cdot \frac{h}{k}$$

Remark. In particular if $u = x - A$ and $v = y - B$ *i.e.*, when $h = 1, k = 1$, we have

$$\bar{x} = A + \frac{\Sigma u}{n}, \quad \bar{y} = B + \frac{\Sigma v}{n},$$

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \quad \text{and} \quad b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2}.$$

Example 8.12. *An investigation into the demand for television sets in 7 towns has resulted in the following data:*

Population x (in thousand)	*11*	*14*	*14*	*17*	*17*	*21*	*25*
No. of T.V. sets demanded, y	*15*	*27*	*27*	*30*	*34*	*38*	*46*

Calculate the regression equation of y on x and estimate the demand for T.V. sets for a town with a population of 30 thousands.

Solution. Computation of Regression Equation of y on x

S. No.	x	y	$u = x - A$, $A = 17$	$v = y - B$, $B = 27$	uv	u^2
1	11	15	−6	−12	72	36
2	14	27	−3	0	0	9
3	14	27	−3	0	0	9
4	17	30	0	3	0	0
5	17	34	0	7	0	0
6	21	38	4	11	44	16
7	25	46	8	19	152	64
$n = 7$			$\Sigma u = 0$	$\Sigma v = 28$	$\Sigma uv = 268$	$\Sigma u^2 = 134$

Regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

We have $$\bar{x} = A + \frac{\Sigma u}{n} = 17 + \frac{0}{7} = 17$$

$$\bar{y} = B + \frac{\Sigma v}{n} = 27 + \frac{28}{7} = 31$$

NOTES

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{7(268) - (0)(28)}{7(134) - (0)^2} = \frac{268}{134} = 2$$

$\therefore$ The required equation is

$$y - 31 = 2(x - 17) \quad \text{or} \quad y = 2x - 3.$$

When population is 30 thousand *i.e.*, $x = 30$, the estimated value of

$$y = 2(30) - 3 = 57.$$

$\therefore$ The estimated demand for T.V. sets is **57.**

Example 8.13. *Obtain the two regression equations from the following data:*

x	25	28	35	32	31	36	29	38	34	32
y	43	46	49	41	36	32	31	30	33	39

Also find the value of y when x is equal to 30.

Solution. Computation of Regression Equations

S. No.	x	y	$u = x - A$, $A = 32$	$v = y - B$, $B = 38$	uv	u^2	v^2
1	25	43	−7	5	−35	49	25
2	28	46	−4	8	−32	16	64
3	35	49	3	11	33	9	121
4	32	41	0	3	0	0	9
5	31	36	−1	−2	2	1	4
6	36	32	4	−6	−24	16	36
7	29	31	−3	−7	21	9	49
8	38	30	6	−8	−48	36	64
9	34	33	2	−5	−10	4	25
10	32	39	0	1	0	0	1
$n = 10$			$\Sigma u = 0$	$\Sigma v = 0$,	$\Sigma uv = -93$	$\Sigma u^2 = 140$	$\Sigma v^2 = 398$

Regression equation of 'y on x'

The regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$$\bar{x} = A + \frac{\Sigma u}{n} = 32 + \frac{0}{10} = 32$$

$$\bar{y} = B + \frac{\Sigma v}{n} = 38 + \frac{0}{10} = 38$$

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{10(-93) - (0)(0)}{10(140) - (0)^2} = -\frac{93}{140} = -0.6643.$$

$\therefore$ The required equation is $y - 38 = -0.6643\,(x - 32)$.

$\Rightarrow$ $$y = -0.6643x + 38 + (0.6643)(32)$$

$\Rightarrow$ $$\mathbf{y = -0.6643x + 59.2576.}$$

When $x = 30$, the estimated value of $y = (-0.6643)\,(30) + 59.2576 = \mathbf{39.3286.}$

Regression equation of 'x on y'

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

$$\bar{x} = 32, \ \bar{y} = 38$$

$$b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} = \frac{10(-93) - (0)(0)}{10(398) - (0)^2} = -\frac{93}{398} = -0.2337.$$

NOTES

$\therefore$ The required equation is $x - 32 = -0.2337\ (y - 38)$.

$\Rightarrow$ $x = -0.2337\ y + 32 + (0.2337)(38)$

$\Rightarrow$ $\mathbf{x = -0.2337y + 40.8806.}$

Example 8.14. *Following are the heights of fathers and sons in inches:*

Height of father	65	66	67	68	69	71	73	67
Height of son	67	68	64	72	70	69	70	68

Find the two lines of regression and estimate the height of the son when the height of the father is 67.5 inches.

Solution. Let the variables 'height of father' and 'height of son' be denoted by x and y respectively.

Computation of Regression Equations

S. No.	x	y	$u = x - A$, $A = 68$	$v = y - B$, $B = 68$	uv	u^2	v^2
1	65	67	-3	-1	3	9	1
2	66	68	-2	0	0	4	0
3	67	64	-1	-4	4	1	16
4	68	72	0	4	0	0	16
5	69	70	1	2	2	1	4
6	71	69	3	1	3	9	1
7	73	70	5	2	10	25	4
8	67	68	-1	0	0	1	0
$n = 8$			$\Sigma u = 2$	$\Sigma v = 4$	$\Sigma uv = 22$	$\Sigma u^2 = 50$	$\Sigma v^2 = 42$

Regression equation of 'y on x'

The regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$$\bar{x} = A + \frac{\Sigma u}{n} = 68 + \frac{2}{8} = 68.25 \text{ inches}$$

$$\bar{y} = B + \frac{\Sigma v}{n} = 68 + \frac{4}{8} = 68.5 \text{ inches}$$

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{8(22) - (2)(4)}{8(50) - (2)^2} = \frac{168}{396} = 0.4242$$

$\therefore$ The required equation is

$y - 68.5 = 0.4242\ (x - 68.25)$.

$\Rightarrow$ $y = 0.4242x + 68.5 - (0.4242)(68.25)$

$\Rightarrow$ $\mathbf{y = 0.4242x + 39.4835.}$

Regression equation of 'x on y'

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

$\bar{x} = 68.25$ inches, $\bar{y} = 68.5$ inches

$$b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} = \frac{8(22) - (2)(4)}{8(42) - (4)^2} = \frac{168}{320} = 0.525$$

$\therefore$ The required equation is

$x - 68.25 = 0.525\ (y - 68.5)$.

$\Rightarrow$ $x = 0.525y + 68.25 - (0.525)(68.5)$

$\Rightarrow$ $\mathbf{x = 0.525y + 32.2875.}$

To find the estimated value of height of son (y) for a given value of height of father (x), we require regression equation of y on x *i.e.*,

$$y = 0.4242x + 39.4835.$$

$\therefore$ When $x = 67.5$ inches, the estimated value of

$$y = (0.4242)(67.5) + 39.4835 = \mathbf{68.117 \text{ inches.}}$$

Example 8.15. *Students of a class have obtained marks as given below in Paper I and Paper II of Statistics:*

Paper I	45	55	56	58	60	65	68	70	75	80	85
Paper II	56	50	48	60	62	64	65	70	74	82	90

Find the means, coefficient of correlation, regression coefficients and regression equations.

Solution. Let the variables 'marks in Paper I' and 'marks in Paper II' be denoted by x and y respectively.

Computation of $\bar{x}, \bar{y}$, r

S. No.	x	y	$u = x - A$, $A = 60$	$v = y - B$, $B = 70$	uv	u^2	v^2
1	45	56	– 15	– 14	210	225	196
2	55	50	– 5	– 20	100	25	400
3	56	48	– 4	– 22	88	16	484
4	58	60	– 2	– 10	20	4	100
5	60	62	0	– 8	0	0	64
6	65	64	5	– 6	– 30	25	36
7	68	65	8	– 5	– 40	64	25
8	70	70	10	0	0	100	0
9	75	74	15	4	60	225	16
10	80	82	20	12	240	400	144
11	85	90	25	20	500	625	400
$n = 11$			$\Sigma u = 57$	$\Sigma v = -49$	$\Sigma uv = 1148$	$\Sigma u^2 = 1709$	$\Sigma v^2 = 1865$

Means $\bar{x} = A + \dfrac{\Sigma u}{n} = 60 + \dfrac{57}{11} = \mathbf{65.1818.}$

$$\bar{y} = B + \frac{\Sigma v}{n} = 70 + \frac{(-49)}{11} = \mathbf{65.5455.}$$

Coefficient of correlation

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{11(1148) - (57)(-49)}{\sqrt{11(1709) - (57)^2}\sqrt{11(1865) - (-49)^2}}$$

$$= \frac{15421}{\sqrt{15550}\sqrt{18114}} = \frac{\mathbf{15421}}{\mathbf{124.70 \times 134.59}} = \mathbf{0.9188.}$$

NOTES

NOTES

Regression coefficients

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{11(1148) - (57)(-49)}{11(1709) - (57)^2} = \frac{15421}{15550} = 0.9917$$

$$b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} = \frac{11(1148) - (57)(-49)}{11(1865) - (-49)^2} = \frac{15421}{18114} = 0.8513.$$

Regression equations

Regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$\Rightarrow$ $y - 65.5455 = 0.9917(x - 65.1818)$

$\Rightarrow$ $y = 0.9917x + 65.5455 - (0.9917)(65.1818)$

$\Rightarrow$ $\mathbf{y = 0.9917x + 0.9047.}$

Regression equation of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$.

$\Rightarrow$ $x - 65.1818 = 0.8513(y - 65.5455)$

$\Rightarrow$ $x = 0.8513y + 65.1818 - (0.8513)(65.5455)$

$\Rightarrow$ $\mathbf{x = 0.8513y + 9.3829.}$

EXERCISE 8.2

1. The following data relates to 'advertising expenditure' (in lakhs of rupees) and 'sales' (in crores of rupees)

Advertising expenditure (in lakhs of rupees)	10	12	15	23	20
Sales (in crores of rupees)	14	17	23	25	21

Estimate (*i*) the sales corresponding to advertising expenditure of ₹ 30 lakhs and (*ii*) the advertising expenditure for a sales target of 35 crores.

2. Find two lines of regression from the following data:

Age of husband	25	22	28	26	35	20	22	40	20	18
Age of wife	18	15	20	17	22	14	16	21	15	14

Hence estimate (*i*) the age of husband when the age of wife is 19 years and (*ii*) the age of wife when the age of husband is 30 years.

3. Find the regression equation of y on x for the following data:

x	78	89	97	69	59	79	68	61
y	125	137	156	112	107	136	124	108

4. Find the two regression equations for the following series. What is the most likely value of x when $y = 20$ and most likely value of y when $x = 22$?

x	35	25	29	31	27	24	33	36
y	23	27	26	21	24	20	29	30

NOTES

5. Find the regression equations for the following data:

x	23	26	39	31	36	21	30	39
y	45	48	45	42	31	39	38	32

Also find:

(*i*) the value of y when $x = 30$

(*ii*) correlation coefficient between x and y.

6. Obtain the lines of regression and show them on the graph paper for the following data:

x	65	66	67	67	68	69	71	71
y	67	68	64	68	70	70	69	68

Answers

1. (*i*) 29.9666 crore rupees (*ii*) 31.75 lakh rupees
2. If x and y respectively represent 'age of husband' and 'age of wife' then the regression equations are $x = 2.23y - 12.76$ and $y = 0.385x + 7.34$

   (*i*) 30 years nearly (*ii*) 19 years nearly
3. $y = 1.212x + 34.725$
4. Regression equation of x on y: $x = 0.543y + 16.425$

   Regression equation of y on x: $y = 0.352x + 14.44$

   $x = 27.285$ when $y = 20$, $y = 22.184$ when $x = 22$
5. Regression equation of y on x: $y = 0.4x + 27.75$

   Regression equation of x on y: $x = 0.511y + 10.185$

   (*i*) 39.75 (*ii*) 0.452
6. Regression equation of y on x: $y = 0.2353x + 52$

   Regression equation of x on y: $x = 0.4615y + 36.618$.

8.8. REGRESSION LINES FOR GROUPED DATA

In case of grouped data if either x or y or both variables represent classes, then their respective mid-points are taken as their representatives.

In this case, if $u = \dfrac{x - A}{h}$, $v = \dfrac{y - B}{k}$,

then the regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$,

where $\bar{x} = A + \left(\dfrac{\Sigma fu}{N}\right)h$,

$$\bar{y} = B + \left(\frac{\Sigma fv}{N}\right)k$$

and $$b_{yx} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fu^2 - (\Sigma fu)^2} \cdot \frac{k}{h}$$

NOTES

The regression line of x on y is

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

where $\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h$,

$$\bar{y} = B + \left(\frac{\Sigma fv}{N}\right)k$$

and $$b_{xy} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fv^2 - (\Sigma fv)^2} \cdot \frac{h}{k}$$

Example 8.16. *Compute regression lines corresponding to the marks obtained by 25 students in Economics and Statistics:*

	Marks in Statistics			
Marks in Economics	30–40	40–50	50–60	60–70
30–40	3	1	1	0
40–50	2	6	1	2
50–60	1	2	2	1
60–70	0	1	1	1

Solution. Let x and y denote the variables 'marks in Statistics' and 'marks in Economics' respectively.

Class of x	:	30–40	40–50	50–60	60–70
Mid-point (x)	:	35	45	55	65
Deviation from A = 45	:	– 10	0	10	20
Step deviation by h = 10 $\left(u = \frac{x - 45}{10}\right)$	:	–1	0	1	2
Class of y	:	30–40	40–50	50–60	60–70
Mid-point (y)	:	35	45	55	65
Deviation from B = 45	:	–10	0	10	20
Step deviation by k = 10 $\left(v = \frac{y - 45}{10}\right)$	:	–1	0	1	2

NOTES

Regression Table

y \ x	v \ u	30–40	40–50	50–60	60–70	f	fv	fv^2	fuv
		−1	0	1	2				
30–40	−1	3 / 3	0 / 1	−1 / 1	0	5	−5	5	2
40–50	0	0 / 2	0 / 6	0 / 1	0 / 2	11	0	0	0
50–60	1	−1 / 1	0 / 2	2 / 2	2 / 1	6	6	6	3
60–70	2	0	0 / 1	2 / 1	4 / 1	3	6	12	6
f		6	10	5	4	N = 25	$\Sigma fv = 7$	$\Sigma fv^2 = 23$	$\Sigma fuv = 11$
fu		−6	0	5	8	$\Sigma fu = 7$			
fu^2		6	0	5	16	$\Sigma fu^2 = 27$			
fuv		2	0	3	6	$\Sigma fuv = 11$			

Now $$\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h = 45 + \left(\frac{7}{25}\right)10 = 47.8$$

$$\bar{y} = B + \left(\frac{\Sigma fv}{N}\right)k = 45 + \left(\frac{7}{25}\right)10 = 47.8$$

$$b_{yx} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fu^2 - (\Sigma fu)^2} \cdot \frac{k}{h} = \frac{25(11) - (7)(7)}{25(27) - (7)^2} \cdot \frac{10}{10} = \frac{2260}{6260} = 0.361$$

$$b_{xy} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fv^2 - (\Sigma fv)^2} \cdot \frac{h}{k} = \frac{25(11) - (7)(7)}{25(23) - (7)^2} \cdot \frac{10}{10} = \frac{2260}{5260} = 0.429$$

The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

or $$y - 47.8 = 0.361\,(x - 47.8)$$

or $$y = 0.361x + 47.8 - (0.361)(47.8)$$

or $$\mathbf{y = 0.361x + 30.544.}$$

The regression line of x on y is $x - \bar{x} = b_{yx}(y - \bar{y})$

or $$x - 47.8 = 0.429\,(y - 47.5)$$

or $$x = 0.429y + 47.8 - (0.429)(47.8)$$

or $$\mathbf{x = 0.429y + 27.294.}$$

NOTES

EXERCISE 8.3

1. The following table gives the frequency, according to groups of marks obtained by 67 students in an intelligence test. Compute the regression lines between the variables age (x) and marks (y):

Test marks	*Age (in years)*			
	18	*19*	*20*	*21*
200—250	4	4	2	1
250—300	3	5	4	2
300—350	2	6	8	5
350—400	1	4	6	10

2. Following is the distribution of students according to their height and weight:

Height (in inches)	*Weight (in lbs)*			
	90—100	*100—110*	*110—120*	*120—130*
50—55	4	7	5	2
55—60	6	10	7	4
60—65	6	12	10	7
65—70	3	8	6	3

Obtain (*i*) the coefficients of regression and (*ii*) the regression equations.

Answers

1. $y = 21.5134x - 109.7157$, $x = 0.008y + 17.1727$
2. (*i*) $b_{yx} = 0.152$, $b_{xy} = 0.041$, (*ii*) $y = 0.152x + 99.93$, $x = 0.041y + 55.88$.

8.9. PROPERTIES OF REGRESSION COEFFICIENTS AND REGRESSION LINES

(*i*) We have $b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$.

σ_x and σ_y are always non-negative.

$\therefore$ The signs of b_{yx} and b_{xy} are same as that of r.

$\therefore$ The signs of regression coefficients and correlation coefficient are same.

Thus b_{yx}, b_{xy} and r are all either positive or negative.

(*ii*) $b_{yx} \cdot b_{xy} = r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y} = r^2$.

Now $0 \le r^2 \le 1$ because $-1 \le r \le 1$.

$\therefore \quad 0 \le b_{yx} b_{xy} \le 1$.

$\therefore$ **The product of regression coefficients is non-negative and cannot exceed one.**

NOTES

(*iii*) $b_{yx} \cdot b_{xy} = r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y} = r^2$

$\therefore \quad \mathbf{r = \pm \sqrt{b_{yx}\, b_{xy}}}$

The sign of r is taken as that of regression coefficients.

(*iv*) The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$\Rightarrow \quad y = b_{yx}\, x + (\bar{y} - b_{yx}\, \bar{x})$

$\therefore$ **When y is kept on the left side, then the coefficient of x on the right side gives the regression coefficient of y on x.**

For example, let $4x + 7y - 9 = 0$ be the regression line of y on x.

We write this as $y = -\frac{4}{7}x + \frac{9}{7}$.

$\therefore$ Regression coefficient of y on x = coefficient of $x = -\frac{4}{7}$

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

$\Rightarrow \quad x = b_{xy}\, y + (\bar{x} - b_{xy}\, \bar{y})$

$\therefore$ **When x is kept on the left side, then the coefficient of y on the right side gives the regression coefficient of x on y.**

For example, let $5x + 9y - 8 = 0$ be the regression line of x on y.

We write this as $x = -\frac{9}{5}y + \frac{8}{5}$.

$\therefore$ Regression coefficient of x on y = coefficient of $y = -\frac{9}{5}$

(*v*) The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

This equation is satisfied by the point $(\bar{x}, \bar{y})$. This point also lies on the regression line of x on y:

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

$\therefore$ **The point $(\bar{x}, \bar{y})$ is common to both regression lines. In other words, if the correlation between the variables is not perfect, then the regression lines intersect at $(\bar{x}, \bar{y})$.**

(*vi*) **Angle between the lines of regression**

The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$\Rightarrow \quad y = b_{yx}\, x + (\bar{y} - b_{yx}\, \bar{x}) \quad \therefore$ Slope $= b_{yx} = m_1$ (say)

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

$\Rightarrow \quad y = \frac{1}{b_{xy}}x + \left(\bar{y} - \frac{1}{b_{xy}}\bar{x}\right) \quad \therefore$ Slope $= \frac{1}{b_{xy}} = m_2$ (say)

Let θ be the acute angle between the regression lines.

$$\therefore \quad \tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right| = \left|\frac{b_{yx} - \frac{1}{b_{xy}}}{1 + b_{yx} \cdot \frac{1}{b_{xy}}}\right| = \left|\frac{b_{yx} b_{xy} - 1}{b_{xy} + b_{yx}}\right|$$

NOTES

$$= \left| \frac{r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y} - 1}{r\frac{\sigma_x}{\sigma_y} + r\frac{\sigma_y}{\sigma_x}} \right| = \left| \frac{r^2 - 1}{r\left(\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x \sigma_y}\right)} \right|$$

$$= \frac{|r^2 - 1||\sigma_x \sigma_y|}{|r||\sigma_x^2 + \sigma_y^2|} = \frac{(1 - r^2)\sigma_x \sigma_y}{|r|(\sigma_x^2 + \sigma_y^2)}$$

$$\therefore \quad \tan\theta = \frac{(1-r^2)\sigma_x\sigma_y}{|r|(\sigma_x^2 + \sigma_y^2)}$$

Particular cases:

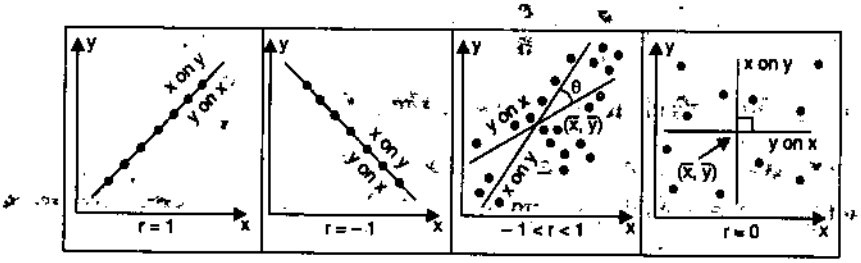
(*i*) **r = 0.** In this case, tan θ is not defined.

∴ θ = 90° *i.e.*, the regression lines are *perpendicular* to each other.

(*ii*) **r = 1 (or – 1).** In this case, tan θ = 0.

∴ The regression lines are *coincident*, because the point $(\bar{x}, \bar{y})$ is on both the regression lines.

Thus, we see that if the variables are not correlated, then the regression lines are perpendicular to each other and if the variables are perfectly correlated, then the regression lines are coincident. The closeness of regression lines measure the degree of linear correlation between the variables.



Example 8.17. *Find the mean of the variable X and Y and correlation coefficient from the following informations:*

Regression equation of Y on X: $2Y - X - 50 = 0$.

Regression equation of X on Y: $3Y - 2X - 10 = 0$.

Solution. Regression equation of Y on X is $2Y - X - 50 = 0$ ...(1)

Regression equation of X on Y is $3Y - 2X - 10 = 0$ ...(2)

Multiplying (1) by 2, we get $4Y - 2X - 100 = 0$ ...(3)

Subtracting (3) from (2), we get $-Y + 90 = 0 \Rightarrow Y = 90$

∴ (1) $\Rightarrow$ $2(90) - X - 50 = 0 \Rightarrow X = 130$

$$\bar{X} = 130, \quad \bar{Y} = 90.$$

(1) $\Rightarrow$ $Y = \frac{1}{2}X + 25 \Rightarrow b_{yx} = \frac{1}{2}$

(2) $\Rightarrow$ $X = \frac{3}{2}Y - 5 \Rightarrow b_{xy} = \frac{3}{2}$

Now $r = +\sqrt{b_{yx} \cdot b_{xy}}$ (Regression coefficients are + ve)

$$= \sqrt{\frac{1}{2} \times \frac{3}{2}} = \frac{\sqrt{3}}{2} = 0.866.$$

NOTES

Example 8.18. *Out of the following two regression lines, find the line of regression of y on x:*

$$4x - 5y + 33 = 0 \quad \text{and} \quad 20x - 9y - 107 = 0.$$

Solution. The regression lines are

$$4x - 5y + 33 = 0 \qquad ...(1)$$

$$20x - 9y - 107 = 0. \qquad ...(2)$$

Let (1) be the regression line of y on x.

$\therefore$ (2) is the regression line of x on y.

(1) $\Rightarrow$ $\quad y = \frac{4}{5}x + \frac{33}{5} \quad \therefore \quad b_{yx} = \frac{4}{5}$

(2) $\Rightarrow$ $\quad x = \frac{9}{20}y + \frac{107}{20} \quad \therefore \quad b_{xy} = \frac{9}{20}$

b_{yx} and b_{xy} are of same signs.

Also $\quad b_{yx}\, b_{xy} = \frac{4}{5} \times \frac{9}{20} = \frac{36}{100} \leq 1.$

$\therefore$ Our choice of regression lines is correct.

$\therefore$ Regression line of y on x is $\mathbf{4x - 5y + 33 = 0.}$

Example 8.19. *Find the regression coefficients b_{yx} and b_{xy} when the lines of regression are*

$$4x + 3y + 7 = 0 \quad \text{and} \quad 3x + 4y + 8 = 0.$$

Solution. The regression lines are

$$4x + 3y + 7 = 0 \qquad ...(1) \qquad \text{and} \qquad 3x + 4y + 8 = 0 \qquad ...(2)$$

Let (1) be the regression line of y on x.

$\therefore$ (2) is the regression line of x on y.

(1) $\Rightarrow$ $\quad y = -\frac{4}{3}x - \frac{7}{3} \quad \therefore \quad b_{yx} = -\frac{4}{3}$

(2) $\Rightarrow$ $\quad x = -\frac{4}{3}y - \frac{8}{3} \quad \therefore \quad b_{xy} = -\frac{4}{3}$

$$b_{yx} \cdot b_{xy} = \left(-\frac{4}{3}\right)\left(-\frac{4}{3}\right) = \frac{16}{9} > 1$$

This is impossible, because $0 \leq b_{yx} \cdot b_{xy} \leq 1$.

$\therefore$ Our supposition is wrong.

(1) is the regression line of x on y and (2) is the regression line y on x.

(1) $\Rightarrow$ $\quad x = -\frac{3}{4}y - \frac{7}{4}$

and (2) $\Rightarrow$ $\quad y = -\frac{3}{4}x - 2$

$\therefore \quad b_{xy} = -\frac{3}{4}$ and $b_{yx} = -\frac{3}{4}$.

Example 8.20. *The equations of regression lines are given by $4x - 5y + 35 = 0$ and $5x - 2y = 20$.*

Also, variance of x-series is 9. Find :

(i) mean values of x and y variables

(ii) correlation coefficient between x and y variables

(iii) standard deviation of y series.

NOTES

Solution. The equations of regression lines are

$$4x - 5y + 35 = 0 \quad ...(1)$$

$$5x - 2y - 20 = 0 \quad ...(2)$$

(*i*) (1) × 2 $\Rightarrow$ $5x - 10y + 70 = 0$ ...(3)

(2) × 5 $\Rightarrow$ $25x - 10y - 100 = 0$ ...(4)

(3) – (4) $\Rightarrow$ $-17x + 170 = 0 \Rightarrow x = 10$

$\therefore$ (1) $\Rightarrow$ $4(10) - 5y + 35 = 0$

$\Rightarrow$ $5y = 40 + 35 = 75 \Rightarrow y = 15$

$\therefore$ (10, 15) is on both lines.

$\therefore$ $\bar{x} = 10, \bar{y} = 15.$

(*ii*) Let (1) be the regression line of y on x.

$\therefore$ (2) is the regression line of x on y.

(1) $\Rightarrow$ $y = \frac{4}{5}x + 7$ $\therefore$ $b_{yx} = \frac{4}{5}$

(2) $\Rightarrow$ $x = \frac{2}{5}y + 4$ $\therefore$ $b_{xy} = \frac{2}{5}$

$\therefore$ b_{yx} and b_{xy} are of same sign.

Also $b_{yx} \cdot b_{xy} = \left(\frac{4}{5}\right)\left(\frac{2}{5}\right) = \frac{8}{25} \leq 1$

$\therefore$ Our choice of regression lines is correct.

$\therefore$ $b_{yx} = \frac{4}{5}$ and $b_{xy} = \frac{2}{5}$

$\therefore$ $r = +\sqrt{b_{yx} \cdot b_{xy}} = +\sqrt{\frac{4}{5} \times \frac{2}{5}} = \frac{2\sqrt{2}}{5} = 0.5657.$

(*iii*) $b_{yx} = \frac{4}{5}$ $\Rightarrow$ $r\frac{\sigma_y}{\sigma_x} = \frac{4}{5}$

$\Rightarrow$ $\frac{2\sqrt{2}}{5} \cdot \frac{\sigma_y}{3} = \frac{4}{5}$ $\quad (\because \text{ var. } x = 9 \Rightarrow \sigma_x = \sqrt{9} = 3)$

$\Rightarrow$ $\sigma_y = \frac{4 \times 3}{2\sqrt{2}} = 2\sqrt{2} = 4.2426.$

Example 8.21. *Equations of two regression lines are:*

$$4x + 3y + 7 = 0 \quad \text{and} \quad 3x + 4y + 8 = 0$$

Find:

(*i*) *mean of x, mean of y;*

(*ii*) *regression coefficients b_{yx} and b_{xy} and*

(*iii*) *correlation coefficient between x and y.*

Solution. The equations of regression lines are:

$$4x + 3y + 7 = 0 \quad ...(1)$$

and $3x + 4y + 8 = 0$ ...(2)

(*i*) (1) × 4 $\Rightarrow$ $16x + 12y + 28 = 0$ ...(3)

(2) × 3 $\Rightarrow$ $9x + 12y + 24 = 0$ ...(4)

(3) – (4) $\Rightarrow$ $7x + 4 = 0 \Rightarrow x = -4/7.$

NOTES

$\therefore$ (1) $\Rightarrow$ $4(-4/7) + 3y + 7 = 0$

$\Rightarrow$ $y = \dfrac{-7 + (16/7)}{3} = -\dfrac{11}{7}$

$\therefore$ $(-4/7, -11/7)$ is on both lines.

Since regressions lines intersects at $(\bar{x}, \bar{y})$, we have

$\bar{x} = -4/7, \bar{y} = -11/7.$

(*ii*) Let (1) be the regression line of y on x.

$\therefore$ (2) is the regression line of x on y.

(1) $\Rightarrow$ $y = -\dfrac{4}{3}x - \dfrac{7}{3}$ $\quad\therefore\quad b_{yx} = -\dfrac{4}{3}$

(2) $\Rightarrow$ $x = -\dfrac{4}{3}y - \dfrac{8}{3}$ $\quad\therefore\quad b_{xy} = -\dfrac{4}{3}$

$\therefore$ $b_{yx} \cdot b_{xy} = \left(-\dfrac{4}{3}\right)\left(-\dfrac{4}{3}\right) = \dfrac{16}{9} > 1$

This is impossible, because $0 \le b_{yx} \cdot b_{xy} \le 1$.

$\therefore$ Our supposition is wrong.

$\therefore$ (1) is the regression line of x on y and (2) is the regression line of y on x.

(1) $\Rightarrow$ $x = -\dfrac{3}{4}y - \dfrac{7}{4}$. $\quad\therefore\quad b_{xy} = -\dfrac{3}{4}$

(2) $\Rightarrow$ $y = -\dfrac{3}{4}x - 2$ $\quad\therefore\quad b_{yx} = -\dfrac{3}{4}$.

(*iii*) $r = -\sqrt{b_{yx} b_{xy}} = -\sqrt{\left(-\dfrac{3}{4}\right)\left(-\dfrac{3}{4}\right)} = -\dfrac{3}{4}$.

EXERCISE 8.4

1. In a problem of regression analysis, the two regression coefficients are found to be -0.6 and -1.4. What is the correlation coefficient?
2. Out of the following two regression lines, find the regression line of x on y:
   (*i*) $13x - 10y + 11 = 0, 2x - y - 1 = 0$
   (*ii*) $x + 4y + 11 = 0, 4x + y - 7 = 0$.
3. Find the correlation coefficient when the lines of regression are
   $2x - 9y + 6 = 0, x - 2y + 1 = 0$.
4. The equations of two regression lines obtained in a correlation analysis are as follows:
   $3x + 13y = 19, x + 3y = 5$.
   Obtain (*i*) the means of x and y
   (*ii*) the regression coefficients b_{yx} and b_{xy}
   (*iii*) the correlation coefficient.
5. In a partially destroyed record, the following data are available:
   Variance of $x = 25$.
   Regression equation of x on y: $5x - y = 22$.
   Regression equation of y on x: $64x - 45y = 24$.
   Find:
   (*i*) Mean values of x and y.
   (*ii*) Coefficient of correlation between x and y
   (*iii*) Standard deviation of y.

NOTES

6. The regression equations of a bivariate distribution are:
   Regression equation of y on x: $4y = 9x + 15$
   Regression equation of x on y: $25x = 6y + 7$.
   Find:
   (*i*) Coefficient of correlation.
   (*ii*) The ratio of means of x and y.
   (*iii*) The ratio of S.D. of x and y.
7. In a partially destroyed laboratory record of an analysis of correlation data, the following results are legible:
   Variance of $x = 9$
   Regression equations: $4x - 5y + 33 = 0$
   $20x - 9y = 107$.
   On the basis of above information, find:
   (*i*) the mean values of x and y.
   (*ii*) standard deviation of y-series, and
   (*iii*) the coefficient of correlation.
8. The equations of regression lines are given to be $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$. Find the values of $\bar{x}$, $\bar{y}$ and r.

Answers

1. -0.9165
2. (*i*) $2x - y - 1 = 0$ (*ii*) $4x + y - 7 = 0$ 3. $r = 0.6667$
4. (*i*) $\bar{x} = 2, \bar{y} = 1$ (*ii*) $b_{yx} = -\frac{3}{13}$, $b_{xy} = -3$ (*iii*) $r = -\frac{3}{\sqrt{13}}$
5. (*i*) $\bar{x} = 6, \bar{y} = 8$ (*ii*) $r = 8/15$ (*iii*) $\sigma_y = 40/3$
6. (*i*) $r = 0.7348$ (*ii*) $\bar{x}/\bar{y} = 59/219$ (*iii*) $\sigma_x/\sigma_y = \sqrt{8}/\sqrt{75}$
7. (*i*) $\bar{x} = 13, \bar{y} = 17$ (*ii*) $\sigma_y = 4$, (*iii*) $r = 0.6$
8. $\bar{x} = 4, \bar{y} = 7, r = -0.5$.

8.10. SUMMARY

- In statistics, *regression analysis* is concerned with the measure of average relationship between variables. Here we shall deal with the derivation of appropriate functional relationships between variables. Regression explains the nature of relationship between variables.
- There are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable (or regressed variable or predicted variable or explained variable).* The variable which influences the value of dependent variable is called *independent variable (or regressor or predictor or explanator).*
- If there are only two variables under consideration, then the regression is called **simple regression.** If there are more than two variables under consideration then the regression is called **multiple regression.** The regression is called **partial regression** if there are more than two variables under consideration and relation between only two variables is established after excluding the effect of other variables. The simple regression is called **linear regression** if the

point on the scatter diagram of variables lies almost along a line otherwise it is termed as **non-linear regression** or **curvilinear regression.**

8.11. REVIEW EXERCISES

NOTES

1. What are regression coefficients? Show that $r^2 = b_{yx} \cdot b_{xy}$.
2. Point out the role of regression analysis in business with the help of few examples.
3. What is regression? Why are there, in general two regression lines? Under what conditions can there be only one regression line?
3. What is regression analysis? Explain its use in business problems with suitable examples.
4. What do you mean by regression coefficients? What are the uses of regression analysis?

NOTES

9. PROBABILITY

STRUCTURE

9.1. INTRODUCTION

The words 'Probability' and 'Chance' are quite familiar to everyone. Many a times, we come across statements like, "Probably it may rain today", "Chances of his visit to the university are very few", "It is possible that he may pass the examination with good marks". In the above statements, the words probably, chance, possible, etc. convey the

NOTES

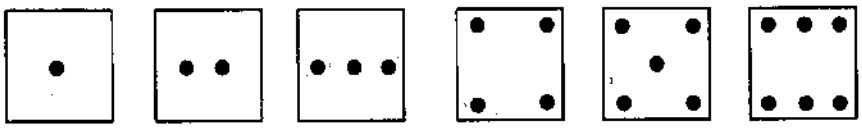
sense of uncertainty about the occurrence of some event. Ordinarily, it may appear that there cannot be any exact measurement for these uncertainties, but in Statistics, we do have methods for calculating the degree of certainty of events in numerical value, provided certain conditions are satisfied.

9.2. RANDOM EXPERIMENT

When we perform experiments in science and engineering, repeatedly under very nearly identical conditions, we get almost the same result. Such experiments are called **deterministic experiments**.

There also exist experiments in which the results may not be essentially the same even if the experiment is performed under very nearly identical conditions. Such experiments are called **random experiments**. If we toss a coin, we may get 'head' or 'tail'. This is a random experiment. Throwing of a die is also a random experiment as any of the six faces of the die may come up. In this experiment, there are six possibilities (1 or 2 or 3 or 4 or 5 or 6).

Remark 1. A *die* is a small cube used in gambling. On its six faces, dots are marked as:



Numbers on a die

Plural of the word die is *dice*. The outcome of throwing a die is the number of dots on its upper most face.

Remark 2. A *pack of cards* consists of four suits called *Spades, Hearts, Diamonds* and *Clubs*. Each suit consists of 13 cards, of which nine cards are numbered from 2 to 10, an ace, a king, a queen and a jack (or knave). Spades and clubs are black faced cards, while hearts and diamonds are red faced cards. The kings, queens and jacks are called *face* cards.

9.3. SAMPLE SPACE

The **sample space** of a random experiment is defined as the set of all possible outcomes of the experiment. The possible outcomes are called **sample points.** The sample space is generally denoted by the letter S. We list the sample space of some random experiments:

Random Experiment	Sample Space
1. Throwing of a fair die	S = {1, 2, 3, 4, 5, 6}
2. Tossing of an unbiased coin	S = {H, T}
3. Tossing of two unbiased coins	S = {HH, HT, TH, TT}
4. A family of two children	S = {BB, BG, GB, GG}

9.4. TREE DIAGRAM

A **Tree diagram** is a device used to enumerate all the logical possibilities of a sequence of steps where each step can occur in a finite number of ways. A tree diagram is constructed from left to right and the number of branches at any point corresponds to the number of ways the next step can occur.

NOTES

Illustration 1. The tree diagram of the sample space of the toss of two coins is shown in the figure.

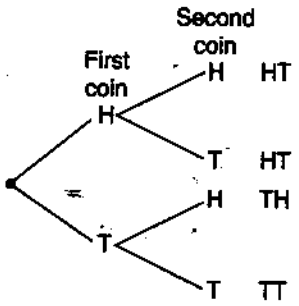
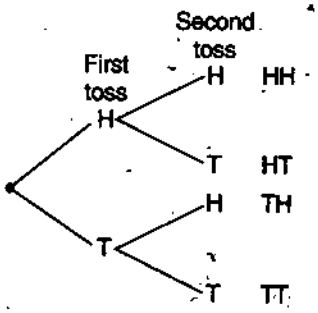


Illustration 2. The tree diagram of the sample space of the two tosses of a coin is shown in the figure.



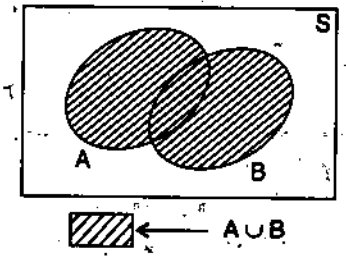
9.5. EVENT

An **event** is defined as a subset of the sample space. An event is called an **elementary (or simple) event** if it contains only one sample point. In the experiment of rolling a die, the event A of getting '3' is a simple event. We write A = {3}. An event is called an **impossible event** if it can never occur. In the above example, the event B = {7} of getting '7' is an impossible event. On the other hand, an event which is sure to occur is called a **certain event.** In the above example of rolling a die, the event C of getting a number less than 7 is a certain event.

In the throwing of two dice, the cases favourable to getting sum 7 are 6 *viz.* (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) and (6, 1).

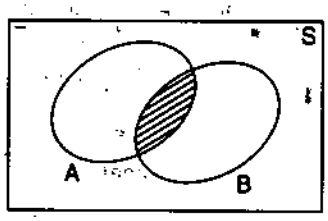
9.6. ALGEBRA OF EVENTS

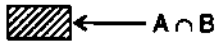
We know that the events of a random experiment are sets, being subsets of the sample space. Thus, we can use set operations to form new events.



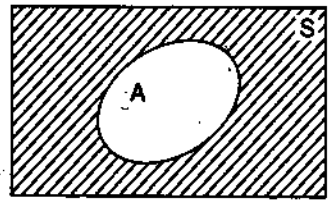
Let A and B be any two events associated with a random experiment.

The event of occurrence of either A or B or both is written as 'A or B' and is denoted by the subset $A \cup B$ of the sample space. The event of occurrence of both A and B is written as 'A and B' and is denoted by the subset $A \cap B$ of the sample space. For simplicity, the event $A \cap B$ is also denoted by 'AB'.





The event of non-occurrence of event A is written as 'not A' and is denoted by the subset A′, which is the complement of set A. The event A′ is called the **complementary event** of the event A.



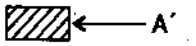


Illustration. Let a die be tossed.

$\therefore$ $S = \{1, 2, 3, 4, 5, 6\}$

Let A = event of getting an even number

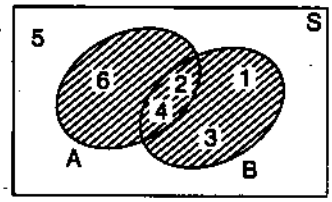
and B = event of getting a number less than 5

$\therefore$ $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$

Here

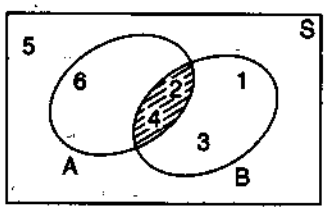
$A \cup B$ = event of occurrence of either A or B or both

$= \{1, 2, 3, 4, 6\}$ (*See figure*)



$A \cap B$ = event of occurrence of both A and B

$= \{2, 4\}$ (*See figure*)



Also, A′ = event of non-occurrence of

$A = \{1, 3, 5\}$

and B′ = event of non-occurrence of

$B = \{5, 6\}$.

9.7. EQUALITY LIKELY OUTCOMES

The outcomes of a random experiment are called **equally likely,** if all of these have equal preferences. In the experiment of tossing a unbiased coin, the outcomes 'Head' and 'Tail' are equally likely.

In our discussion, we shall always assume the outcomes of a random experiment to be equally likely.

9.8. EXHAUSTIVE OUTCOMES

The outcomes of a random experiment are called **exhaustive**, if these cover all the possible outcomes of the experiment. In the experiment of rolling a die, the outcomes 1, 2, 3, 4, 5, 6 are exhaustive.

NOTES

9.9. THREE APPROACHES OF PROBABILITY

There are three approaches of discussing probability of events. These approaches are as follows:

1. Classical approach
2. Empirical approach
3. Subjective approach

We shall first discuss classical approach of probability.

9.10. CLASSICAL APPROACH OF PROBABILITY

Suppose in a random experiment, there are n exhaustive, equally likely outcomes. Let A be an event and there are m outcomes (cases) favourable to the happening of it. Then the **probability** P(A) of the happening of the event A is defined as

$$P(A) = \frac{\text{Total no. of cases favourable to the happening of A}}{\text{Total no. of exhaustive, equally likely cases}} = \frac{m}{n}$$

It may be observed from this definition, that $0 \le m \le n$.

$$\therefore \quad 0 \le \frac{m}{n} \le 1 \quad \text{or} \quad 0 \le P(A) \le 1.$$

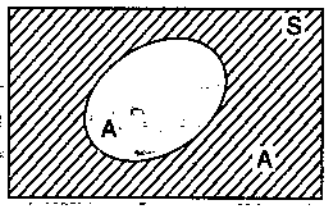
The number of cases favourable to the non-happening of the event A is $n - m$.

$$\therefore \quad P(\text{not A}) = \frac{n-m}{n} = \frac{n}{n} - \frac{m}{n} = 1 - \frac{m}{n} = 1 - P(A).$$

$\therefore$ P(A) + P(not A) = 1. *i.e.* $P(A) + P(\overline{A}) = 1$.

If A is a *sure event*, the $P(A) = \frac{n}{n} = 1$ and if A happen to be an *impossible event*, then

$$P(A) = \frac{0}{n} = 0.$$



9.11. 'ODDS IN FAVOUR' AND 'ODDS AGAINST' AN EVENT

The ratio of cases in favour of A and cases against A is called the **odds in favour** of A. Similarly, the ratio of cases against A and cases in favour of A is called the **odds against** A.

If odds in favour of A are $m : n$, then (*i*) odds against A are $n : m$ and (*ii*) probability of $A = \frac{m}{m+n}$.

If odds against A are $m : n$, then (*i*) odds in favour of A are $n : m$ and (*ii*) probability of $A = \frac{n}{m+n}$.

NOTES

Illustration. Let $P(A) = \frac{3}{5}$.

Let total number of cases be 5λ.

$\therefore$ $P(A) = \frac{3\lambda}{5\lambda}$ and this implies that cases in favour of A are 3λ and cases against A are $5\lambda - 3\lambda = 2\lambda$.

$\therefore$ odds in favour of A are $3\lambda : 2\lambda$ or **3 : 2** and odds against A are $2\lambda : 3\lambda$ or **2 : 3.**

Example 9.1. *Find the probability of getting the sum 10 in a single throw of two dice.*

Solution. Here $S = \{(1, 1), (1, 2), (1, 3), \ldots\ldots, (6, 5), (6, 6)\}$.

$\therefore$ No. of possible outcomes $= 6 \times 6 = 36$.

Let A be the event of getting sum 10.

$\therefore$ $A = \{(4, 6), (5, 5), (6, 4)\}$

$$P(A) = \frac{3}{36} = \frac{1}{12}.$$

Remark. In the above experiment, the sample point (a, b) means that 'a' is on the first die and 'b' is on the second die.

Example 9.2. *Find the probability of getting sum 10 in two throws of a die.*

Solution. Here $S = \{(1, 1), (1, 2), (1, 3), \ldots, (6, 5), (6, 6)\}$

$\therefore$ No. of possible outcomes $= 6 \times 6 = 36$

Let A be the event of getting sum 10

$\therefore$ $A = \{(4, 6), (5, 5), (6, 4)\}$

$$P(A) = \frac{3}{36} = \frac{1}{12}.$$

Remark. In the above experiment, the sample ponit (a, b) means that 'a' occurred in the first toss and 'b' occurred in the second toss.

Example 9.3. *From a bag containing 4 red and 5 green balls, a ball is drawn at random. What is the probability that it is a red ball?*

Solution. Total no. of balls $= 4 + 5 = 9$

No. of red balls $= 4$

$\therefore$ Prob. of getting a red ball $= \frac{\text{Toral no. of red ball}}{\text{Total no. of balls}} = \frac{4}{9}$.

Example 9.4. *Three coins are tossed. Find the probability of getting at least two heads.*

Solution. Let S be the sample space.

$\therefore$ $S = (HHH, HHT, HTH, THH, HTT, THT, TTH, TTT)$

Let A be the event of getting at least two heads.

$\therefore$ $A = \{HHH, HHT, HTH, THH\}$

$\therefore$ P(at least two heads)

$$= P(A) = \frac{\text{No. of cases favourable to A}}{\text{Total no. of exhaustive, equally likely cases}} = \frac{4}{8} = \frac{1}{2}.$$

Example 9.5. *Find the probability of getting a 'King' or a 'Queen' in a single draw from a well-shuffled pack of playing cards.*

Solution. Let A be the event of getting a king or a queen in the draw.

$\therefore$ No. of favourable cases for the happening of the event A $= 4 + 4 = 8$

NOTES

Total no. of cases = 52

$\therefore$ P(King or Queen) = P(A) = $\frac{8}{52} = \frac{2}{13}$.

Example 9.6. *Two unbiased dice are thrown. Find the probability that the total of the numbers on the dice is greater than 8.*

Solution. Let S be the sample space.

$\therefore$ S = {(1, 1), (1, 2), (1, 3), ..., (6, 5), (6, 6)}

There are $6 \times 6 = 36$ exhaustive, equally likely outcomes.

Let A be the event of getting total greater than 8.

$\therefore$ A = {(3, 6), (4, 5), (5, 4), (6, 3), (4, 6), (5, 5), (6, 4), (5, 6), (6, 5), (6, 6)}

$\therefore$ P(sum > 8) = P(A) = $\frac{\text{No. of cases favourable to A}}{\text{Total No. of cases}} = \frac{10}{36} = \frac{5}{18}$.

Example 9.7. *Three unbiased dice are thrown simultaneously. Find the probability of getting (i) sum not greater than 5, (ii) sum at least 15, (iii) sum equal to 8.*

Solution. Let S be the sample space.

$\therefore$ S = {(1, 1, 1), (1, 1, 2), ..., (6, 6, 5), (6, 6, 6)}

There are $6 \times 6 \times 6 = 216$ exhaustive equally likely outcomes.

(*i*) Let A = event that sum is not greater than 5.

$\therefore$ A = {(1, 1, 1), (1, 1, 2) (1, 2, 1), (2, 1, 1), (1, 1, 3),
(1, 2, 2), (1, 3, 1), (2, 1, 2), (2, 2, 1), (3, 1, 1)}

$\therefore$ $P(A) = \frac{n(A)}{n(S)} = \frac{10}{216} = \frac{5}{108}$.

(*ii*) Let B = event that sum is at least 15

$\therefore$ B = {(3, 6, 6), (4, 5, 6), (4, 6, 5), (5, 4, 6), (5, 5, 5),
(5, 6, 4), (6, 3, 6), (6, 4, 5), (6, 5, 4), (6, 6, 3),
(4, 6, 6), (5, 5, 6), (5, 6, 5), (6, 4, 6), (6, 5, 5),
(6, 6, 4), (5, 6, 6), (6, 5, 6), (6, 6, 5), (6, 6, 6)}

$\therefore$ $P(B) = \frac{n(B)}{n(S)} = \frac{20}{216} = \frac{5}{54}$.

(*iii*) Let C = event of getting sum 8

$\therefore$ C = {(1, 1, 6), (1, 2, 5), (1, 3, 4), (1, 4, 3), (1, 5, 2),
(1, 6, 1), (2, 1, 5), (2, 2, 4), (2, 3, 3), (2, 4, 2),
(2, 5, 1), (3, 1, 4), (3, 2, 3), (3, 3, 2), (3, 4, 1),
(4, 1, 3), (4, 2, 2), (4, 3, 1), (5, 1, 2), (5, 2, 1), (6, 1, 1)}

$\therefore$ $P(C) = \frac{n(C)}{n(S)} = \frac{21}{216} = \frac{7}{72}$.

Example 9.8. *Find the probability that in a random arrangement of the letters of the world* **DAUGHTER**, *the letter* **D** *occupies the first place.*

Solution. The word **DAUGHTER** contains 8 letters and all are different.

$\therefore$ Total no. of possible arrangements

$= 8! = 40320$

Let A be the event of getting a word with **D** at the first place.

NOTES

$\therefore$ Favourable cases to the event A

= 1 × no. of ways of arranging 7 letters (except D).

= 1 × 7 ! = 5040

$\therefore$ P(D occupies first place)

$$P(A) = \frac{5040}{40320} = \frac{1}{8}.$$

Example 9.9. *Find the probability that in a random arrangement of letters of the word* **MATHEMATICS,** *the consonants occur together.*

Solution. The word **MATHEMATICS**, contains 2 **M**'s, 2 **A**'s, 2 **T**'s, 1 **H**, 1 **E**, 1 **I**, 1 **C** and 1 **S**.

$\therefore$ Total no. of exhaustive cases

$$= \frac{11!}{2!2!2!} = 4989600$$

For finding the no. of favourable cases to the event under consideration, we shall consider all consonents **M, T, H, M, T, C, S** as one block. So in arranging **2 A's, 1 E's, 1 I's, 1 (MTHMTCS)**, all the consonents will occur together. The consonents **MTHMTCS** can arrange themselves in

$$\frac{7!}{2!2!} = 1260 \text{ ways}$$

$\therefore$ No. of favourable cases $= \frac{5!}{2!1!1!1!} \times 1260 = 75600$

$\therefore$ P(consonants are together) $= \frac{75600}{4989600} = 0.01515.$

Example 9.10. *A bag contains 10 red and 8 black balls. Two balls are drawn at random. Find the probability that:*

(i) both balls are red.

(ii) one ball is red and the other is black.

Solution. (*i*) Total number of balls = 10 + 8 = 18.

(*ii*) P(both red balls) $= \frac{\text{No. of selections of 2 out of 10 red balls}}{\text{No. of selections of 2 out of 18 balls}}$

$$= \frac{{}^{10}C_2}{{}^{18}C_2} = \frac{10 \times 9}{1 \times 2} + \frac{18 \times 17}{1 \times 2} = \frac{10 \times 9}{18 \times 17} = \frac{5}{17}.$$

(*ii*) P(one red ball and one black ball)

$$= \frac{\begin{pmatrix}\text{No. of selections of} \\ \text{1 out of 10 red balls}\end{pmatrix}\begin{pmatrix}\text{No. of selections of} \\ \text{1 out of 8 black balls}\end{pmatrix}}{\text{No. of selections of 2 out of 18 balls.}}$$

$$= \frac{{}^{10}C_1 \times {}^{8}C_1}{{}^{18}C_2}$$

$$= (10 \times 8) + \frac{18 \times 17}{2} = \frac{10 \times 8}{9 \times 17} = \frac{80}{153}.$$

NOTES

Example 9.11. *From a group of 8 children (5 boys and 3 girls) three children are selected at random. Calculate the probabilities that the selected group contains:*

(i) no girl *(ii) only one girl*

(iii) only one particular girl.

Solution. No. of girls = 3

No. of boys = 5

Let A_0, A_1, A_2 and A_3 be the events that there are no girl, one girl, two girls and three girls in the random selection of three children.

(*i*) Required probability $= P(A_0) = \frac{{}^5C_3}{{}^8C_3} = \frac{5}{28}$.

(*ii*) Required probability $= P(A_1) = \frac{{}^3C_1 \times {}^5C_2}{{}^8C_3} = \frac{15}{28}$.

(*iii*) The particular girl can be selected in only one way. The remaining two children in the selection must be boys.

$\therefore$ Required probability $= \frac{1 \times {}^5C_2}{{}^8C_3} = \frac{5}{28}$.

EXERCISE 9.1

1. Find the probability of getting an odd number in a single throw of a fair die.
2. A single letter is selected at random from the word **"PROBABILITY"**. What is the probability that it is a vowel?
3. Find the probability of getting at most two heads in a single toss of three coins.
4. From a pack of 52 playing cards, one card is drawn at random. What is the probability that it will be a queen of club or king of diamond?
5. Two unbiased dice are thrown. Find the probability that the total of the numbers on the dice is 8.
6. Tickets are numbered from 1 to 100. These are well-shuffled and a ticket is drawn at random. What is the probability that the drawn ticket has a number which is greater than 75?
7. The following table gives a distribution of marks:

Marks	*No. of students*
0—10	7
10—20	10
20—30	15
30—40	28
40—50	23
50—60	45
60—70	12
70—80	7
80—90	2
90—100	1

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

(*i*) his marks are below 50 (*ii*) his marks are below 70.

(*iii*) his marks are between 30 and 60 (*iv*) his marks are above 70.

NOTES

8. Find the probability that in a random arrangement of the letters of the word **VOWEL**, the letter **V** occupies the first place.
9. Find the probability that in a random arrangement of letters the word **BHARAT**, the two **A** occupies the first two places.
10. Find the probability that in a random arrangement of the letters of the word **STATISTICS**, the three **T** are in the beginning.
11. Find the probability that in a random arrangement of the letters of the word **STATISTICS**, the three **T** are together.
12. Four cards are drawn from a pack of playing cards. Find the probability that none is a king.
13. A bag contains 6 white, 4 red and 10 black balls. Two balls are drawn at random. Find the probability that both balls are black.
14. A bag contains 7 white, 5 red and 8 black balls. Two balls are drawn at random. Calculate the probability that none is white.

Answers

1. $\frac{1}{2}$ 2. $\frac{4}{11}$ 3. $\frac{7}{8}$ 4. $\frac{1}{26}$
5. $\frac{5}{36}$ 6. $\frac{1}{4}$
7. (*i*) 0.5533 (*ii*) 0.9333 (*iii*) 0.64 (*iv*) 0.0667
8. 0.2 9. 0.0667 10. 0.0083 11. 0.0667
12. $\frac{^{48}C_4}{^{52}C_4} = 0.7187$ 13. $\frac{^{10}C_2}{^{20}C_2} = \frac{9}{38}$ 14. $\frac{39}{95}$

9.12. MUTUALLY EXCLUSIVE EVENTS

Two events associated with a random experiment are said to be **mutually exclusive** if both cannot occur together in the same trial. In the experiment of throwing a die, the events A = {1, 4} and B = {2, 5, 6} are mutually exclusive events. In the same experiment, the events A = {1, 4} and C = {2, 4, 5, 6} are not mutually exclusive because if 4 appear on the die, then it is favourable to both events A and C. The definition of mutually exclusive events can also be extended to more than two events. We say that more than two events are mutually exclusive if the happening of one of these rules out the happening of all other events. The events A = {1, 2}, B = {3} and C = {6}, are mutually exclusive in connection with the experiment of throwing a single die.

9.13. ADDITION THEOREM (FOR MUTUALLY EXCLUSIVE EVENTS)

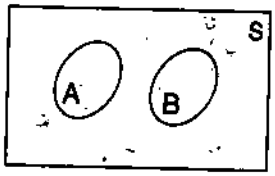
If A and B are two mutually exclusive events associated with a random experiment, then

$$P(A \text{ or } B) = P(A) + P(B).$$

Proof. Let n be the total number of exhaustive, equally like cases of the experiment.

NOTES

Let m_1 and m_2 be the number of cases favourable to the happening of the events A and B respectively.



$$P(A) = \frac{m_1}{n}$$

and $$P(B) = \frac{m_2}{n}$$

Since the events are given to be mutually exclusive, therefore there cannot be any sample point common to both events A and B.

$\therefore$ The event A or B can happen in exactly $m_1 + m_2$ ways.

$$\therefore \quad P(A \text{ or } B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B).$$

Hence, **P(A or B) = P(A) + P(B).**

This theorem can also be extended to even more than two events.

If $A_1, A_2, \ldots\ldots, A_k$ are m.e. events, then

$P(A_1$ or A_2 or ..... or $A_k) = P(A_1) + P(A_2) + \ldots\ldots + P(A_k)$.

Example 9.12. *A box contains 4 red balls, 4 green balls and 7 white balls. What is the probability that a ball drawn is either red or white?*

Solution. Total no. of balls = 4 + 4 + 7 = 15

4 Red
4 Green
7 White

Let A = event of drawing a red ball

B = event of drawing a white ball.

The events A and B are *m.e.* because a ball cannot be both red and white.

$$P(A) = \frac{\text{No. of red balls}}{\text{Total no. of balls}} = \frac{4}{15}$$

$$P(B) = \frac{\text{No. of white balls}}{\text{Total no. of balls}} = \frac{7}{15}$$

Now A or B is the event of drawing either a red ball or a white ball. By addition theorem, the required probability

$$P(A \text{ or } B) = P(A) + P(B) = \frac{4}{15} + \frac{7}{15} = \frac{11}{15}$$

Example 9.13. *In a single throw of 2 dice, determine the probability of getting total 7 or 11.*

Solution. Here S = {(1, 1), (1, 2), ..., (6, 5), (6, 6)}.

Let A and B be the events of getting total 7 and 11 respectively.

$\therefore$ A = {(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)}

and B = {(5, 6), (6, 5)}

The events A and B are mutually exclusive and

$$P(A) = \frac{6}{36} = \frac{1}{6} \quad \text{and} \quad P(B) = \frac{2}{36} = \frac{1}{18}$$

$\therefore$ *By addition theorem,*

$$P(\text{total is 7 or 9}) = P(A \cup B) = P(A) + P(B)$$

$$= \frac{1}{6} + \frac{1}{18} = \frac{3+1}{18} = \frac{2}{9}.$$

NOTES

Example 9.14. *In a single throw of two dice, find the probability that neither a doublet nor a total of 9 will appear.*

Solution. Here $S = \{(1, 1), (1, 2), \ldots\ldots, (6, 5), (6, 6)\}$.

$\therefore$ Number of possible outcomes $= 6 \times 6 = 36$.

Let E = event that doublet is occurred

and F = event that sum is 9.

$\therefore$ $E = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

and $F = \{(3, 6), (4, 5), (5, 4), (6, 3)\}$.

$\therefore$ $P(E) = \frac{6}{36} = \frac{1}{6}$ and $P(F) = \frac{4}{36} = \frac{1}{9}$.

P(neither a doublet nor a total of 9)

$$= P(E^c \cap F^c) = P((E \cup F)^c) = 1 - P(E \cup F) \quad \ldots(1)$$

The events E and F are *m.e.*

$\therefore$ By *addition theorem*,

$$P(E \cup F) = P(E) + P(F) = \frac{1}{6} + \frac{1}{9} = \frac{5}{18}$$

$\therefore$ (1) $\Rightarrow$ $P(E^c \cap F^c) = 1 - \frac{5}{18} = \frac{13}{18}$.

EXERCISE 9.2

1. A and B are mutually exclusive events for which $P(A) = 0.3$, $P(B) = p$ and $P(A \cup B) = 0.5$. Find the value of p.
2. Find the probability that a card drawn from a pack of playing cards is either a 'queen' or a 'king'.
3. A and B are two mutually exclusive events of an experiment.
   If P('not A') = 0.65, $P(A \cup B) = 0.65$ and $P(B) = p$, find the value of p.
   [**Hint.** Use $P(A \cup B) = (1 - P(\text{'not A'})) + P(B)$.]
4. From a set of 17 cards, numbered 1, 2, 3, ..., 16, 17, one is drawn at random. Show that the chance that its number is divisible by 3 or 7 is 7/17.
5. Find the probability of getting the sum 9 or 11 in a single throw of two dice.
6. In a single throw of three dice, find the probability of getting a total of 17 or 18.

Answers

1. 0.2 **2.** $\frac{2}{13}$ **3.** 0.3 **5.** $\frac{1}{6}$

6. $\frac{1}{54}$

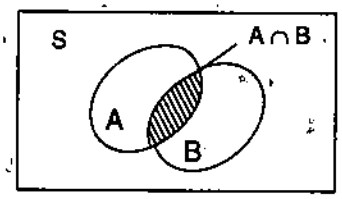
9.14. ADDITION THEOREM (GENERAL)

If A and B are two events not necessarily mutually exclusive, associated with a random experiments, then

$$P(A \text{ or } B) = P(A) + P(B) - P(AB).$$

NOTES

Proof. Let n be the total number of exhaustive equally likely cases of the experiment.



Let m_1 and m_2 be the number of cases favourable to the happening of the events A and B respectively.

$$\therefore \quad P(A) = \frac{m_1}{n} \quad \text{and} \quad P(B) = \frac{m_2}{n}.$$

Since the events are given to be not necessarily non-mutually exclusive, there may be some sample points common to both events A and B.

Let m_3 be number of these common sample points. m_3 will be zero in case A and B are mutually exclusive.

$$\therefore \quad P(AB) = \frac{m_3}{n}.$$

The m_3 sample points which are common to both events A and B, are included in the events A and B separately.

$\therefore$ Number of sample points in the event A or B

$$= m_1 + m_2 - m_3.$$

m_3 is subtracted from $m_1 + m_2$ to avoid counting of common sample points twice.

$$\therefore \quad P(A \text{ or } B) = \frac{m_1 + m_2 - m_3}{n} = \frac{m_1}{n} + \frac{m_2}{n} - \frac{m_3}{n} = P(A) + P(B) - P(AB).$$

Hence, **P(A or B) = P(A) + P(B) – P(AB).**

Corollary 1. If events A and B happen to be mutually exclusive events, then P(AB) = 0 and in this case *addition theorem* implies

$$P(A \text{ or } B) = P(A) + P(B) - P(AB) = P(A) + P(B) - 0 = P(A) + P(B)$$

$$\therefore \quad P(A \text{ or } B) = P(A) + P(B).$$

This is the same as the **addition theorem** for mutually exclusive events.

Corollary 2. If A, B, C are three events associated with a random experiment, then

$$\mathbf{P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(BC) - P(CA) - P(AB) + P(ABC).}$$

Example 9.15. *Find the probability that a card drawn from a pack of playing cards is either a 'queen' or a 'spade'.*

Solution. Total number of cases (cards) = 52

Let A = event of drawing a 'queen'

B = event of drawing a 'spade'

$$\therefore \quad P(A) = \frac{4}{52} \quad \text{and} \quad P(B) = \frac{13}{52}.$$

Here the events are not mutually exclusive as drawing of the card 'queen of spade' is common to both events.

$$\therefore \quad P(AB) = \frac{1}{52}.$$

By *addition theorem*, the probability of getting either a queen or a spade is

$$P(A \text{ or } B) = P(A) + P(B) - P(AB) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}.$$

NOTES

Example 9.16. *One number is drawn from numbers 1 to 150. Find the probability that it is divisible by either 3 or 5.*

Solution. Here $S = \{1, 2, 3, ..., 149, 150\}$.

Let A = event that the number is divisible by 3

$\therefore$ $A = \{3, 6, 9, ..., 147, 150\}$

$\therefore$ $P(A) = \frac{50}{150}$.

Let B = event that the number is divisible by 5.

$\therefore$ $B = \{5, 10, 15 ....., 145, 150\}$

$\therefore$ $P(B) = \frac{30}{150}$.

The events A and B are not *m.e.* because the sample points 15, 30, 45, ..., 150 are common to both.

$\therefore$ $AB = \{15, 30, 45, ..., 135, 150\}$

$\therefore$ $P(AB) = \frac{10}{150}$.

By *addition theorem*, the required probability of getting a multiple of either 3 or 5 is

$$P(A \text{ or } B) = P(A) + P(B) - P(AB) = \frac{50}{150} + \frac{30}{150} - \frac{10}{150} = \frac{70}{150} = \frac{7}{15}.$$

Example 9.17. *A student applies for a job in two firms X and Y. The probability of his being selected in firm X is 0.7 and being rejected in the firm Y is 0.5. The probability of at least one of his application being rejected is 0.6. What is the probability that he will be selected in one of the firms?*

Solution. Let A = event of getting selected in X

B = event of getting selected in Y

$\therefore$ $P(A) = 0.7$, $P(B) = 1 - 0.5 = 0.5$

$P(AB) = 1 - P$ (rejecting in at least one firm) $= 1 - 0.6 = 0.4$

$\therefore$ P(selected in one of the firm)

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

$= 0.7 + 0.5 - 0.4 = \mathbf{0.8}$.

EXERCISE 9.3

1. Find the probability that a card drawn from a pack of playing cards is either a 'king' or a 'club'.
2. A drawer contains 50 bolts and 150 nuts. Half of the bolts and half of the nuts are rusted. If one item is chosen at random, what is the probability that it is rusted or is a bolt ?
3. From 30 tickets marked with first 30 numerals, one is drawn at random. Find the probability that it is:
   (*i*) a multiple of 5 or 7 (*ii*) a multiple of 3 or 7.
4. A construction company is bidding for two contracts A and B. The probability that the company will get contract A is 3/5, the probability that the company will get contract B is 1/3 and the probability that the company will get both the contracts is 1/8. What is the probability that the company will get contract A or B?

NOTES

5. The probability that a contractor will get a plumbing contract is 2/3 and the probability that he will not get an electric contract is 5/9. The probability of getting at least one contract is 4/5. What is the probability that he will get both contracts ?

$\left[\text{Hint. } \frac{4}{5} = \frac{2}{3} + \left(1 - \frac{5}{9}\right) - P(AB).\right]$

6. Find the probability of getting at least one five, in a single throw of two dice.

Answers

1. $\frac{4}{13}$ 2. $\frac{5}{8}$ 3. (*i*) $\frac{1}{3}$ (*ii*) $\frac{13}{30}$
4. $\frac{97}{120}$ 5. $\frac{14}{45}$ 6. $\frac{11}{36}$

9.15. CONDITIONAL PROBABILITY

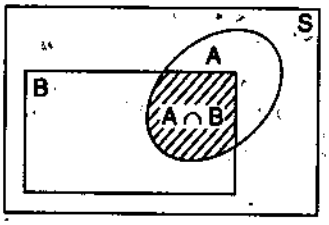
Let us consider the random experiment of throwing a die. Let *A* be the event of getting an odd number on the die.

$$S = \{1, 2, 3, 4, 5, 6\} \quad \text{and} \quad A = \{1, 3, 5\}.$$

$$\therefore \quad P(A) = \frac{3}{6} = \frac{1}{2}$$

Let $B = \{2, 3, 4, 5, 6\}$.

If, after the die is thrown, we are given the information that the event B has occurred, then the probability of event A will no more be 1/2, because in this case, the favourable cases are three and the total number of possible outcomes will be five and not six. The probability of event A, with the condition that event B has happened will be 3/5. This conditional probability is denoted as P(A/B). Let us define the concept of conditional probability in a formal manner.



Let A and B be any two events associated with a random experiment. The probability of occurrence of event A when the event B has already occurred is called the **conditional probability** of A when B is given and is denoted as P(A/B). The conditional probability P(A/B) is meaningful only when $P(B) \neq 0$, *i.e.* when B is not an impossible event.

By definition,

$$P(A/B) = \text{Probability of occurrence of event A when the event B has already occurred.}$$

$$= \frac{\text{no. of cases favourable to B which are also favourable to A}}{\text{no. of cases favourable to B}}$$

$$\therefore \quad \mathbf{P(A/B) = \frac{\text{no. of cases favourable to } A \cap B}{\text{no. of cases favourable to B}}}$$

$$\text{Also, } P(A/B) = \frac{\frac{\text{no. of cases favourable to } A \cap B}{\text{no. of cases in the sample space}}}{\frac{\text{no. of cases favourable to B}}{\text{no. of cases in the sample space}}}$$

$$\therefore \quad \mathbf{P(A/B) = \frac{P(A \cap B)}{P(B)}}.$$

Remark 1. If $P(A) \neq 0$, the $P(B/A) = \dfrac{P(B \cap A)}{P(A)} = \dfrac{P(A \cap B)}{P(A)}$.

Remark 2. If A and B are *m.e.* events, then

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0}{P(B)} = 0 \quad \text{and} \quad P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{0}{P(A)} = 0.$$

NOTES

$\therefore$ **If A and B are m.e. events, then A/B and B/A are impossible events.**

For an illustration, let us consider the random experiment of throwing two coins.

$\therefore$ $S = \{HH, HT, TH, TT\}$.

Let $A = \{HH, HT\}$, $B = \{HH, TH\}$, $C = \{HH, HT, TH\}$ and $D = \{TT\}$.

$\therefore$ $P(A) = \dfrac{2}{4} = \dfrac{1}{2}$, $P(B) = \dfrac{2}{4} = \dfrac{1}{2}$, $P(C) = \dfrac{3}{4}$, $P(D) = \dfrac{1}{4}$.

A/B is the event of getting A with the condition that B has occurred.

$\therefore$ $P(A/B) = \dfrac{n(A \cap B)}{n(B)} = \dfrac{n\{HH\}}{n\{HH, TH\}} = \dfrac{1}{2}$.

Similarly, $P(A/C) = \dfrac{n\{HH, HT\}}{n\{HH, HT, TH\}} = \dfrac{2}{3}$ and $P(B/C) = \dfrac{n\{HH, TH\}}{n\{HH, HT, TH\}} = \dfrac{2}{3}$.

We observe that $P(A/C) \neq P(A)$, $P(B/C) \neq P(B)$.

The events A and D are *m.e.* and we have

$$P(A/D) = \frac{n(A \cap D)}{n(D)} = \frac{0}{1} = 0 \quad \text{and} \quad P(D/A) = \frac{n(D \cap A)}{n(A)} = \frac{0}{2} = 0.$$

Example 9.18. *If P(E) = 0.40, P(F) = 0.35 and P(E ∪ F) = 0.55, find P(E/F).*

Solution. We have $P(E) = 0.40$, $P(F) = 0.35$, $P(E \cup F) = 0.55$.

By *addition theorem*,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

$\therefore$ $0.55 = 0.40 + 0.35 - P(E \cap F)$.

$\therefore$ $P(E \cap F) = 0.75 - 0.55 = 0.20$.

Required probability, $P(E/F) = \dfrac{P(E \cap F)}{P(F)} = \dfrac{0.20}{0.35} = \dfrac{4}{7}$.

Example 9.19. *A coin is tossed twice and the four possible outcomes are assumed to be equally likely. If E is the event "both head and tail have occurred", and F the event: "at most one tail is observed", find P(E), P(F), P(E/F) and P(F/E).*

Solution. We have $S = \{HH, HT, TH, TT\}$

$E = \{HT, TH\}$ and $F = \{HH, HT, TH\}$.

$\therefore$ $E \cap F = \{HT, TH\}$.

$\therefore$ $P(E) = \dfrac{n(E)}{n(S)} = \dfrac{2}{4} = \dfrac{1}{2}$, $P(F) = \dfrac{n(F)}{n(S)} = \dfrac{3}{4}$,

$$P(E/F) = \frac{n(E \cap F)}{n(F)} = \frac{2}{3} \quad \text{and} \quad P(F/E) = \frac{n(E \cap F)}{n(E)} = \frac{2}{2} = 1.$$

Example 9.20. *A die is thrown twice and the sum of the number appearing is observed to be 6. What is the conditional probability that the number 4 has appeared at least once?*

Solution. Let S be the sample space of the experiment.

$\therefore$ $S = \{(1, 1), (1, 2), \ldots\ldots, (6, 5), (6, 6)\}$.

NOTES

Let A = event of getting sum 6

and B = event of getting 4 at least once.

$\therefore$ A = {(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)}

and B = {(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (1, 4), (2, 4), (3, 4), (5, 4), (6, 4)}.

$\therefore$ $P(A) = \frac{5}{36}$ and $P(B) = \frac{11}{36}$.

Also $A \cap B = \{(4, 2), (2, 4)\}$ $\therefore$ $P(A \cap B) = \frac{2}{36}$

Now, required probability

= Probability of getting 4 on at least one die given that sum in 6

$$= P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{2/36}{5/36} = \frac{2}{5}.$$

Alternatively,

$$P(B/A) = \frac{\text{no. of cases favourable to 'A} \cap \text{B'}}{\text{no. of cases favourable to A}} = \frac{2}{5}.$$

Remark 1. In practical problems, it would be easier to use the formulae:

(*i*) $P(A/B) = \frac{\text{no. of favourable to 'A} \cap \text{B'}}{\text{no. of cases favourable to B}}$

(*ii*) $P(B/A) = \frac{\text{no. of cases favourable to 'B} \cap \text{A' } i.e. \text{ 'A'} \cap \text{B'}}{\text{no. of cases favourable to A}}$

Remark 2. The event $A \cap B$ is same as $B \cap A$ and each consists of sample points which are common to both A and B.

Example 9.21. *One card in drawn from a well shuffled pack of 52 cards. If E is the event "the card drawn is either a king or an ace" and F is the event "the card drawn is either an ace or a jack", then find the probability of the conditional event E/F.*

Solution. There are 4 kings and 4 aces in the pack.

$\therefore$ $P(E) = \frac{4+4}{52} = \frac{2}{13}$

There are 4 aces and 4 jacks in the pack.

$\therefore$ $P(F) = \frac{4+4}{52} = \frac{2}{13}$

The event $E \cap F$ contain 4 aces.

$\therefore$ $P(E \cap F) = \frac{4}{52} = \frac{1}{13}$

$\therefore$ Required probability = $P(E/F) = \frac{P(E \cap F)}{P(F)} = \frac{1/13}{2/13} = \frac{1}{2}$.

Example 9.22. *Three fair coins are tossed. Find the probability that they are all tails, if:*

(i) at least one of the coins show tail *(ii) two coins show tail*

(iii) at least two coins show head *(iv) at most one coin show head.*

Solution. Here S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

Let A = event of getting all tails $\therefore$ A = {TTT}

(*i*) Let B = event that at least one of the coins show tail

$\therefore$ B = {HHT, HTH, THH, HTT, THT, TTH, TTT}

NOTES

$\therefore$ $A \cap B = \{TTT\}$

$\therefore$ Required probability $= P(A/B) = \dfrac{n(A \cap B)}{n(B)} = \dfrac{1}{7}$.

(*ii*) Let B = event that two coins show tail

$\therefore$ $B = \{HTT, THT, TTH\}$

$\therefore$ $A \cap B = \phi$

$\therefore$ Required probability $= \dfrac{n(A \cap B)}{n(B)} = \dfrac{0}{3} = 0.$

(*iii*) Let B = event that at least two coins show head

$\therefore$ $B = \{HHH, HHT, HTH, THH\}$

$\therefore$ $A \cap B = \phi$

$\therefore$ Required probability $= P(A/B) = \dfrac{n(A \cap B)}{n(B)} = \dfrac{0}{4} = 0.$

(*iv*) Let B = event that at most one coin show head.

$\therefore$ $B = \{HTT, THT, TTH, TTT\}$

$\therefore$ $A \cap B = \{TTT\}$

$\therefore$ Required probability $= P(A/B) = \dfrac{n(A \cap B)}{n(B)} = \dfrac{1}{4}$.

Remark. The value of $P(A/B)$ is equal to $\dfrac{n(A \cap B)}{n(B)}$ which is also equal to $\dfrac{P(A \cap B)}{P(B)}$.

EXERCISE 9.4

1. If P (not A) = 0.7, P(B) = 0.7 and P(B/A) = 0.5, then find P(A/B) and $P(A \cup B)$.
2. For two events A and B, P(A) = 0.5, P(B) = 0.6 and $P(A \cap B) = 0.8$. Find the conditional probabilities P (A/B) and P(B/A).
3. A die is thrown. Find that probability that the number obtained is greater than 2 if:
   (*i*) no other information is given, (*ii*) it is given that the number obtained is less than 5.
4. A pair of fair dice is thrown. Find the probability that the sum is 10 or greater if 5 appears on the first die.
5. A pair of fair dice is thrown. If the two numbers appearing are different, find the probability that the sum is 4 or less.
6. The probability that a person stopping at a petrol pump will ask to have his tyres checked is 0.12, the probability that he will ask to have his oil checked is 0.29 and the probability that he will ask to have both of them checked is 0.07.
   (*i*) What is the probability that a person who has oil checked will also have tyre checked?
   (*ii*) What is the probability that a person stopping at the petrol pump will have either tyres or oil checked?

   [**Hint:** Let A and B be the events of getting 'tyres checked' and 'oil checked' respectively.

   $\therefore$ $P(A) = 0.12$, $P(B) = 0.29$, $P(A \cap B) = 0.07$.

   (*ii*) Required probability $= P(A/B) = \dfrac{P(A \cap B)}{P(B)}$

   (*iii*) Required probability $= P(A \cup B) = P(A) + P(B) - P(A \cap B)$.]

Answers

1. 0.2143, 0.85 2. 0.5, 0.6 3. (i) $\frac{2}{3}$ (ii) $\frac{1}{2}$

NOTES

4. $\frac{1}{3}$ 5. $\frac{2}{15}$ 6. (i) $\frac{7}{29}$ (ii) $\frac{17}{50}$

9.16. INDEPENDENT EVENTS

Let A and B be two events associated with a random experiment. We know that

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

$$\therefore \quad P(A \cap B) = P(A)\, P(B/A).$$

In general P(B/A) may or may not be equal to P(B). When P(B/A) and P(B) are equal, then the events A and B are of special importance.

Two events associated with a random experiment are said to be **independent events** if the occurrence or non-occurrence of one event does not affect the probability of the occurrence of the other event. For example, the events A and B are independent events when P(A/B) = P(A) and P(B/A) = P(B).

Theorem. Let A and B be events associated with a random experiment. The events A and B are independent if and only if

$$\mathbf{P(A \cap B) = P(A)\, P(B).}$$

Proof. Let A and B be independent events.

$$\therefore \quad P(A \cap B) = \frac{P(A \cap B)}{P(B)} \times P(B) = P(A/B)\, P(B)$$

$$= P(A)\, P(B) \qquad [\because \quad P(A/B) = P(A)]$$

$$\therefore \quad P(A \cap B) = P(A)\, P(B).$$

Conversely, let $P(A \cap B) = P(A)\, P(B)$.

$$\therefore \quad P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\, P(B)}{P(B)} = P(A)$$

and

$$P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{P(A)\, P(B)}{P(A)} = P(B).$$

$$\therefore \quad P(A/B) = P(A) \quad \text{and} \quad P(B/A) = P(B).$$

$\therefore$ A and B are independent events.

Remark 1. $P(A \cap B) = P(A)\, P(B)$ is the necessary and sufficient condition for the events A and B to be independent.

Remark 2. Let A and B be events associated with a random experiment.

(i) Let A and B be *m.e.* $\therefore \quad P(A \cap B) = 0$

$\therefore \quad P(A \cap B) \neq P(A)\, P(B)$ *i.e.* A and B are not independent events.

$\therefore$ **Mutually exclusive events cannot be independent.**

(ii) Let A and B be independent.

$\therefore \quad P(A \cap B) = P(A)\, P(B)$ *i.e.* $P(A \cap B) \neq 0$.

$\therefore$ A and B are not *m.e.* events.

$\therefore$ **Independent events cannot be mutually exclusive.**

Important observation. If A and B be any two events associated with a random experiment, then their physical description is not sufficient to decide if A and B are independent events or not. A and B are declared to be independent events only when we have P(AB) = P(A) P(B).

9.17. DEPENDENT EVENTS

NOTES

Let A and B be two events associated with a random experiment. If A and B are not independent events, then these are called **dependent events.**

$\therefore$ In case of dependent events, we have $\mathbf{P(A \cap B) = P(A)\ P(B/A).}$

Theorem. Let A and B be events associated with a random experiment. If A and B are independent, then show that the events (*i*) $\overline{A}$, B (*ii*) A, $\overline{B}$ (*iii*) $\overline{A}$, $\overline{B}$ are also independent.

Proof. The events A and B are independent.

$\therefore$ $P(A \cap B) = P(A)\ P(B)$ ...(1)

(*i*) $(A \cap B) \cap (\overline{A} \cap B) = (A \cap \overline{A}) \cap (B \cap B) = \phi \cap B = \phi$

and $(A \cap B) \cup (\overline{A} \cap B) = (A \cup \overline{A}) \cap B = S \cap B = B.$

$\therefore$ The events $A \cap B$ and $\overline{A} \cap B$ are *m.e.* and their union in B.

$\therefore$ By *addition theorem*, we have

$$P(B) = P(A \cap B) + P(\overline{A} \cap B).$$

$$\Rightarrow \quad P(\overline{A} \cap B) = P(B) - P(A \cap B) = P(B) - P(A)\ P(B) \quad \text{[Using (1)]}$$

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

$\therefore$ $\mathbf{P(\overline{A} \cap B) = P(\overline{A})\ P(B)}$ *i.e.* $\overline{A}$ and B are independent.

(*ii*) $(A \cap B) \cap (A \cap \overline{B}) = (A \cap A) \cap (B \cap \overline{B}) = A \cap \phi = \phi$

and $(A \cap B) \cup (A \cap \overline{B}) = A \cap (B \cup \overline{B}) = A \cap S = A$

$\therefore$ The events $A \cap B$ and $A \cap \overline{B}$ are *m.e.* and their union is A.

$\therefore$ By *addition theorem*, we have

$$P(A) = P(A \cap B) + P(A \cap \overline{B}).$$

$$\Rightarrow \quad P(A \cap \overline{B}) = P(A) - P(A \cap B) = P(A) - P(A)\ P(B) \quad \text{[Using (1)]}$$

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

$\therefore$ $\mathbf{P(A \cap \overline{B}) = P(A)\ P(\overline{B})}$ *i.e.* A and $\overline{B}$ are independent.

(*iii*) $(\overline{A} \cap B) \cap (\overline{A} \cap \overline{B}) = (\overline{A} \cap \overline{A}) \cap (B \cap \overline{B}) = \overline{A} \cap \phi = \phi$

and $(\overline{A} \cap B) \cap (\overline{A} \cap \overline{B}) = \overline{A} \cap (B \cup \overline{B}) = \overline{A} \cap S = \overline{A}.$

$\therefore$ The events $\overline{A} \cap B$ and $\overline{A} \cap \overline{B}$ are *m.e.* and their union is $\overline{A}$.

$\therefore$ By *addition theorem*, we have

$$P(\overline{A}) = P(\overline{A} \cap B) + P(\overline{A} \cap \overline{B}) \quad \text{...(1)}$$

$$\Rightarrow \quad P(\overline{A} \cap \overline{B}) = P(\overline{A}) - P(\overline{A} \cap B) = P(\overline{A}) - P(\overline{A})\ P(B)$$

[Using part (*i*)]

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

$\therefore$ $\mathbf{P(\overline{A} \cap \overline{B}) = P(\overline{A})\ P(\overline{B})}$ *i.e.* $\overline{A}$ and $\overline{B}$ are independent.

NOTES

Example 9.23. *If A and B are independent events such that P(A ∪ B) = 0.6 and P(A) = 0.2, find P(B).*

Solution. We have P(A ∪ B) = 0.6 and P (A) = 0.2.

By *addition theorem*, we have

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

$$\therefore \quad P(A \cup B) = P(A) + P(B) - P(A)\,P(B) \quad (\because \text{ A and B are independent})$$

$$\Rightarrow \quad 0.6 = 0.2 + P(B) - (0.2)\,P(B)$$

$$\Rightarrow \quad 0.4 = P(B)(1 - 0.2)$$

$$\Rightarrow \quad (0.8)\,P(B) = 0.4 \quad \Rightarrow \quad P(B) = \frac{0.4}{0.8} = \frac{1}{2} = 0.5.$$

Example 9.24. *A coin is tossed thrice and all the eight outcomes are assumed equally likely. In which of the following cases are the events E and F independent?*

(i) E : the first throw results in head.

F : the last throw results in tail.

(ii) E : the number of heads is two.

F : the last throw results in head.

(iii) E : the number of heads is odd.

F : the number of tails is odd.

Solution. Let S be the sample space.

$$\therefore \quad S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

(*i*) Here E = {HHH, HHT, HTH, HTT) and F = {HHT, HTT, THT, TTT}.

$$\therefore \quad P(E) = \frac{4}{8} = \frac{1}{2} \quad \text{and} \quad P(F) = \frac{4}{8} = \frac{1}{2}.$$

There are 2 cases favourable to the event E ∩ F, namely HHT and HTT.

$$\therefore \quad P(E \cap F) = \frac{2}{8} = \frac{1}{4}.$$

Also, $P(E)P(F) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = P(E \cap F)$.

∴ The events E and F are **independent.**

(*ii*) Here E = {HHT, HTH, THH} and F = {HHH, HTH, THH, TTH}.

$$\therefore \quad P(E) = \frac{3}{8} \quad \text{and} \quad P(F) = \frac{4}{8} = \frac{1}{2}.$$

There are 2 cases favourable to the event E ∩ F, namely HTH and THH.

$$\therefore \quad P(E \cap F) = \frac{2}{8} = \frac{1}{4}.$$

Also, $P(E)\,P(F) = \frac{3}{8} \times \frac{1}{2} = \frac{3}{16} \neq P(E \cap F)$.

∴ The events E and F are **not independent.**

(*iii*) Here E = {HHH, HTT, THT, TTH} and F = {HHT, HTH, THH, TTT}.

$$\therefore \quad P(E) = \frac{4}{8} = \frac{1}{2} \quad \text{and} \quad P(F) = \frac{4}{8} = \frac{1}{2}.$$

There is no case favourable to the event E ∩ F.

NOTES

$\therefore \quad P(E \cap F) = \frac{0}{8} = 0.$

Also, $P(E)P(F) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \neq P(E \cap F).$

$\therefore$ The events E and F are **not independent.**

EXERCISE 9.5

1. (*i*) Two events A and B are such that P(A) = 0.6, P(B) = 0.2 and P(A ∩ B) = 0.8. Does this imply that A and B are independent?
   (*ii*) The events A and B are given to be independent. Find P(B) if it is given that P(A) = 0.35 and P(A ∪ B) = 0.60.
2. (*i*) If P(not B) = 0.65, P(A ∪ B) = 0.85 and A and B are independent events, find P(A).
   (*ii*) If P(not A) = 0.4, P(A ∪ B) = 0.75 and A, B are given to be independent events, find the value of P(B).
3. A coin is tossed twice and all possible outcomes are assumed to be equally likely. A is the event : both head and tail have occurred and B is the event : at least one tail has occurred". Show that A and B are not independent.
4. One card is drawn from a pack of 52 cards so that each card is equally likely to be selected. A is the event : "the card is a heart" and B is the event : "the card is a king." Show that A and B are independent.
5. A die is thrown and the 6 possible outcomes are assumed to be equally likely. If E is the event : "the number appearing is a multiple of 3", and F is the event : "the number appearing is even". Show that the events E and F are independent.

Answers

1. (*i*) No. (*ii*) $\frac{5}{13}$ 2. (*i*) $\frac{10}{13}$ (*ii*) $\frac{3}{8}$

9.18. INDEPENDENT EXPERIMENTS

Two random experiments are said to be **independent experiments** if the occurrence or non-occurrence of an event in one experiment does not in any way affect the probability of occurrence of any event in the other experiment. For example, two tosses of a coin are independent experiments.

9.19. MULTIPLICATION THEOREM

If A and B be events associated with independent experiments E_1 and E_2 respectively, then prove that

$$P(AB) = P(A)P(B).$$

Proof. Since the random experiments E_1 and E_2 are independent, the sample spaces of the experiments are not affected by the events.

Let n_1 and n_2 be the numbers of exhaustive, equally likely cases in the first and second experiment respectively.

Let m_1 be the number of cases favourable to the happening of the event A out of n_1 cases of the first experiment.

NOTES

$$\therefore \quad P(A) = \frac{m_1}{n_1}$$

Let m_2 be the number of cases favourable to the happening of the event B out of n_2 cases of the second experiment.

$$\therefore \quad P(B) = \frac{m_2}{n_2}$$

By the **Fundamental principle of events**, the number of cases favourable to the happening of the event AB in this specified order is $m_1 m_2$. Also the number of extaustive, equally likely cases in the combined experiment is $n_1 n_2$.

$$\therefore \quad P(AB) = \frac{m_1 m_2}{n_1 n_2} = \frac{m_1}{n_1} \cdot \frac{m_2}{n_2} = P(A)P(B).$$

$$\therefore \quad \mathbf{P(AB) = P(A)\ P(B).}$$

The theorem can also be extended to even more than two events.

Let $A_1, A_2, \ldots\ldots, A_k$ be k events associated with random experiments $E_1, E_2, \ldots\ldots, E_k$ with probabilities $p_1, p_2, \ldots\ldots, p_k$ respectively, then

$$P(A_1 A_2 \ldots\ldots, A_k) = P(A_1)\ P(A_2) \ldots\ldots P(A_k) = p_1\ p_2 \ldots\ldots\ p_k.$$

Also $\quad P(\text{not } A_1) = P(\overline{A}_1) = 1 - p_1$

$$P(\text{not } A_2) = P(\overline{A}_2) = 1 - p_2$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$P(\text{not } A_k) = P(\overline{A}_k) = 1 - p_k.$$

Since $A_1, A_2, \ldots\ldots A_k$ are associated with independents experiment, therefore, the events $\overline{A}_1, \overline{A}_2, \ldots\ldots, \overline{A}_k$ are also associated with independent experiments.

Now P(event of happening of at least one of $A_1, A_2, \ldots\ldots, A_k$)

$$= 1 - P(\text{event of happening of none of } A_1, A_2, \ldots\ldots, A_k)$$

$$= 1 - P(\overline{A}_1 \overline{A}_2 \ldots\ldots \overline{A}_k) = 1 - P(\overline{A}_1)\ P(\overline{A}_2) \ldots\ldots P(\overline{A}_k) \quad \textit{(By Multiplication theorem)}$$

$$= 1 - (1 - p_1)(1 - p_2) \ldots\ldots (1 - p_k).$$

Remark. If A and B are events associated with experiments which are not independent, then the probability of the event 'AB' is found by using the result:

$$\mathbf{P(AB) = P(A)\ P(B/A).}$$

This result can also be extended to more than two experiments.

Example 9.25. *A and B appeared for an interview for two posts. Probability of A's rejection is 2/5 and that of B's selection is 4/7. Find the probability that one of them is selected.*

Solution. The random experiments 'interview of A' and 'interview of B' are experiment.

Let $\quad E$ = event that A is selected

and $\quad F$ = event that B is selected.

$$\therefore \quad P(\overline{E}) = \frac{2}{5} \quad \text{and} \quad P(F) = \frac{4}{7}.$$

Also, $P(E) = 1 - P(\overline{E}) = 1 - \frac{2}{5} = \frac{3}{5}$

and $P(\overline{F}) = 1 - P(F) = 1 - \frac{4}{7} = \frac{3}{7}$

NOTES

Required probability = P(only one is selected)

$= P(E\overline{F} \cup \overline{E}F) = P(E\overline{F}) + P(\overline{E}F)$

(Using addition theorem)

$= P(E)\,P(\overline{F}) + P(\overline{E})\,P(F)$

(Using multiplication theorem)

$= \frac{3}{5} \times \frac{3}{7} + \frac{2}{5} \times \frac{4}{7} = \frac{17}{35}$.

Example 9.26. *The odds in favour of one student passing a test are 3 : 7. The odds against another student passing it are 3 : 5. What is the probability that both pass the test?*

Solution. Let A = event that first pass the test.

$\therefore \quad P(A) = \frac{3}{3+7} = \frac{3}{10}$

Let B = event the second pass the test.

$\therefore \quad P(B) = \frac{5}{3+5} = \frac{5}{8}$

The random experiments of results of students are independent.

$\therefore$ P(both pass the test) = $P(AB) = P(A)P(B) = \frac{3}{10} \times \frac{5}{8} = \frac{3}{16}$.

Example 9.27. *A speaks truth in 60% of the cases and B in 90% of the cases. In what percentage of cases, are they likely to contradict each other in stating the same fact?*

Solution. The random experiments of speeches of *A* and *B* are independent.

Let E = event of A speaking truth

and F = event of B speaking truth.

$\therefore \quad P(E) = \frac{60}{100} = \frac{6}{10}$ and $P(F) = \frac{90}{100} = \frac{9}{10}$.

Probability of A and B contradicting each other = $P(E\overline{F}$ or $\overline{E}F)$

$= P(E\overline{F}) + P(\overline{E}F) = P(E)\,P(\overline{F}) + P(\overline{E})\,P(F)$

$= P(E)(1 - P(F)) + (1 - P(E))P(F)$

$= \frac{6}{10}\left(1 - \frac{9}{10}\right) + \left(1 - \frac{6}{10}\right)\frac{9}{10} = \frac{6}{10} \times \frac{1}{10} + \frac{4}{10} \times \frac{9}{10} = \frac{42}{100}$.

$\therefore$ A and B are likely to contradicts each other in **42%** cases.

Example 9.28. *A problem in statistics is given to five students A, B, C, D and E. Their chances of solving the problem are $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}$ and $\frac{1}{6}$ respectively. What is the probability that the problem will be solved?*

Solution. The random experiments of trying the problem by the given students are independent.

NOTES

Let A_1 = event that A fails to solve the problem
A_2 = event that B fails to solve the problem
A_3 = event that C fails to solve the problem
A_4 = event that D fails to solve the problem
A_5 = event that E fails to solve the problem.

$\therefore \quad P(A_1) = 1 - \frac{1}{2} = \frac{1}{2} \qquad P(A_2) = 1 - \frac{1}{3} = \frac{2}{3}$

$P(A_3) = 1 - \frac{1}{4} = \frac{3}{4} \qquad P(A_4) = 1 - \frac{1}{5} = \frac{4}{5}$

$P(A_5) = 1 - \frac{1}{6} = \frac{5}{6}$

Now, P(event that the problem is solved by at least one student)

= 1 – P(event that the problem is not solved by any of the five students)

$= 1 - P(A_1 A_2 A_3 A_4 A_5)$

$= 1 - P(A_1) P(A_2) P(A_3) P(A_4) P(A_5)$ (By Multiplication Theorem)

$$= 1 - \left(\frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \times \frac{4}{5} \times \frac{5}{6}\right) = 1 - \frac{1}{6} = \frac{5}{6}.$$

Example 9.29. *Three bags contains 7 white 8 red, 9 white 6 red and 5 white 7 red balls respectively. One ball, at random, is drawn from each bag. Find the probability that all of them are of the same colour.*

Solution. The three random experiments of drawing balls from given bags are independent.

Let W_i and R_i be the events of drawing white ball and red ball respectively from the ith bag, $i = 1, 2, 3$.

Required probability = P(all balls of same colour)

$= P(W_1W_2W_3 \text{ or } R_1R_2R_3)$

$= P(W_1 W_2 W_3) + P(R_1R_2R_3)$

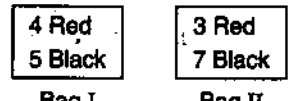
$= P(W_1)P(W_2)P(W_3) + P(R_1)P(R_2)P(R_3)$

$$= \frac{7}{7+8} \times \frac{9}{9+6} \times \frac{5}{5+7} + \frac{8}{7+8} \times \frac{6}{9+6} \times \frac{7}{5+7}$$

$$= \frac{7}{15} \times \frac{9}{15} \times \frac{5}{12} + \frac{8}{15} \times \frac{6}{15} \times \frac{7}{12} = \frac{651}{2700} = \frac{217}{900}.$$

Example 9.30. *A bag has 4 red and 5 black balls, a second bag has 3 red and 7 black balls. One ball is drawn from the first and two from the second. Find the probability that out of three balls, two are black and one is red.*

Solution. The random experiments of 'drawing one ball from first bag' and 'drawing two balls from second bag' are independent.



We are to find the probability two black balls and one red ball.

NOTES

∴ Required probability

$$= P(F_r, S_{bb} \text{ or } F_b, S_{rb}),$$

where F_r is the event of drawing red ball from the first bag, etc.

$$= P(F_r S_{bb}) + P(F_b S_{rb}) = P(F_r)\, P(S_{bb}) + P(F_b)\, P(S_{rb})$$

$$= \frac{4}{4+5} \times \frac{{}^7C_2}{{}^{3+7}C_2} + \frac{5}{4+5} \times \frac{{}^3C_1\, {}^7C_1}{{}^{3+7}C_2}$$

$$= \frac{4}{9} \times \frac{\frac{7\times 6}{1\times 2}}{\frac{10\times 9}{1\times 2}} + \frac{5}{9} \times \frac{3\times 7}{\frac{10\times 9}{1\times 2}} = \frac{4}{9} \times \frac{7}{15} + \frac{5}{9} \times \frac{7}{15} = \frac{7}{15}.$$

Example 9.31. *In a hockey match, the probability of winning of Indian team against Pakistani team in 1/4. Three matches are played. Find the probability that:*

(i) India looses all the matches.

(ii) India wins at least one match.

(iii) India wins two matches.

Solution. The random experiments of matches between Indian team and Pakistani team are independent.

Let A_i to the event of winning of Indian team in the ith match, $i = 1, 2, 3$.

$$\therefore \qquad P(A_i) = \frac{1}{4} \quad \text{and} \quad P(\overline{A}_i) = 1 - \frac{1}{4} = \frac{3}{4}$$

(*i*) P(India loosing all the matches)

$$= P(\overline{A}_1\, \overline{A}_2\, \overline{A}_3) = P(\overline{A}_1)\, P(\overline{A}_2)\, P(\overline{A}_3) = \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}.$$

(*ii*) P(India winning at least one match)

$$= 1 - P(\text{India loosing all the matches})$$

$$= 1 - \frac{27}{64} = \frac{37}{64}. \qquad \textbf{[Using part (i)]}$$

(*iii*) P(India winning two matches)

$$= P(A_1\, A_2\, \overline{A}_3 \text{ or } A_1\, \overline{A}_2\, A_3 \text{ or } \overline{A}_1\, A_2\, A_3)$$

$$= P(A_1\, A_2\, \overline{A}_3) + P(A_1\, \overline{A}_2\, A_3) + P(\overline{A}_1\, A_2\, A_3)$$

$$= P(A_1)\, P(A_2)\, P(\overline{A}_3) + P(A_1)\, P(\overline{A}_2)\, P(A_3) + P(\overline{A}_1)\, P(A_2)\, P(A_3)$$

$$= \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} + \frac{1}{4} \times \frac{3}{4} \times \frac{1}{4} + \frac{3}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{9}{64}.$$

Example 9.32. *Three groups of children consists of respectively 3 girls and 1 boy, 2 girls and 2 boys, 1 girl and 3 boys. One child is selected at random from each group. Find the chance that three selected children comprise 1 girl and 2 boys.*

Solution. The random experiments of drawing one child from each group are independent.

3 Girls 1 Boy	2 Girls 2 Boys	1 Girl 3 Boys
Group I	Group II	Group III

NOTES

Let G_i and B_i be the events of selecting a girl and a boy respectively from the ith group, $i = 1, 2, 3$.

$\therefore$ P(1 girl and 2 boys)

$$= P(G_1B_2B_3 \text{ or } B_1G_2B_3 \text{ or } B_1B_2G_3)$$
$$= P(G_1B_2B_3) + P(B_1G_2B_3) + P(B_1B_2G_3)$$
$$= P(G_1)P(B_2)P(B_3) + P(B_1)P(G_2)P(B_3) + P(B_1)P(B_2)P(G_3)$$
$$= \left(\frac{3}{4}\times\frac{2}{4}\times\frac{3}{4}\right)+\left(\frac{1}{4}\times\frac{2}{4}\times\frac{3}{4}\right)+\left(\frac{1}{4}\times\frac{2}{4}\times\frac{1}{4}\right) = \frac{18+6+2}{64} = \frac{13}{12}.$$

Example 9.33. *A bag contains 6 white ball and 4 black balls. Two balls are drawn at random one by one without replacement. Find the probability that both balls are white.*

Solution. Since the first ball is not replaced before the second draw, the random experiments of drawing balls are not independent.

Let W_1 = event that first ball is white

W_2 = event that second ball is white

P(both balls are white)

$$= P(W_1W_2) = P(W_1)P(W_2/W_1)$$
$$= \frac{6}{6+4}\times\frac{6-1}{(6-1)+4} = \frac{6}{10}\times\frac{5}{9} = \frac{1}{3}.$$

Example 9.34. *From a pack of playing cards, two cards are drawn one by one without replacement. Find the probability that:*

(i) first is king and second is queen

(ii) one is king and other in queen.

Solution. Since the first card is not replaced before the second draw, the random experiments of drawing cards are not independent.

Let K_i and Q_i be the events of drawing a king and a queen respectively in the ith draw, $i = 1, 2$.

(i) P(first is king and second is queen)

$$= P(K_1Q_2) = P(K_1)\,P(Q_2/K_1)$$
$$= \frac{4}{52}\times\frac{4}{51} = \frac{4}{663}.$$

(ii) P(one king and one queen)

$$= P(K_1Q_2 \text{ or } Q_1K_2) = P(K_1\,Q_2) + P(Q_1\,K_2)$$
$$= P(K_1)\,P(Q_2/K_1) + P(Q_1)\,P(K_2/Q_1)$$
$$= \frac{4}{52}\times\frac{4}{51} + \frac{4}{52}\times\frac{4}{51} = \frac{8}{663}.$$

Example 9.35. *From a well-shuffled pack of playing cards, two cards are drawn at random one by one. Find the probability that they are both kings if the first card is : (i) replaced, (ii) not replaced before the second draw.*

Solution. Let K_1 and K_2 be the events of getting kings in the first draw and second draw respectively before the second draw.

(i) Since the first card is replaced, the random experiments are independent.

$\therefore$ P(both kings) $= P(K_1K_2) = P(K_1)\,P(K_2) = \frac{4}{52}\times\frac{4}{52} = \frac{1}{169}$

(*ii*) Since the first card is not replaced, the random experiments are not independent.

$$\therefore \text{P(both kings)} = P(K_1 K_2) = P(K_1)\, P(K_2/K_1) = \frac{4}{52} \times \frac{3}{51} = \frac{1}{221}.$$

Example 9.36. *A bag contains 8 red and 5 white balls. Two successive drawings of three balls are made such that (i) balls are replaced before second trial, (ii) balls are not replaced before second trial. Find the probability that 1st drawing will give 3 white and the 2nd 3 red balls.*

Solution. Let W_1 and R_2 be the events of getting 3 white balls in the first draw and 3 red balls in the second draw respectively.

8 Red
5 White

(*i*) Since the balls of first draw are replaced, the random experiments are independent.

$$\therefore \quad P(W_1 R_2) = P(W_1)\, P(R_2) = \frac{{}^5C_3}{{}^{13}C_3} \times \frac{{}^8C_3}{{}^{13}C_3} = \frac{10}{286} \times \frac{56}{286} = 0.0068$$

(*ii*) Since the balls of first draw are not replaced, the random experiments are not independent.

$$\therefore \quad P(W_1 R_2) = P(W_1)\, P(R_2/W_1) = \frac{{}^5C_3}{{}^{13}C_3} \times \frac{{}^8C_3}{{}^{10}C_3} = \frac{10}{286} \times \frac{56}{120} = 0.0163.$$

Example 9.37. *In each of a set of games, it is 2 to 1 in favour of the winner of the previous game. What is the chance that the player who wins the first game shall win at least three of the next four games?*

Solution. Let W_i be the event that the winner of the first game wins the ith game, $i = 2, 3, 4, 5$.

$\therefore$ P(Winner of the first game wins at least 3 out of the next 4 games)

$$= P(W_2 W_3 W_4 \overline{W}_5 \text{ or } W_2 W_3 \overline{W}_4 W_5 \text{ or } W_2 \overline{W}_3 W_4 W_5 \text{ or } \overline{W}_2 W_3 W_4 W_5 \text{ or } W_2 W_3 W_4 W_5)$$

$$= P(W_2 W_3 W_4 \overline{W}_5) + P(W_2 W_3 \overline{W}_4 W_5) + P(W_2 \overline{W}_3 W_4 W_5) + P(\overline{W}_2 W_3 W_4 W_5) + P(W_2 W_2 W_3 W_4)$$

$$= P(W_2)\, P(W_3/W_2)\, P(W_4/W_2 W_3)\, P(\overline{W}_5/W_2 W_3 W_4) + P(W_2)\, P(W_3/W_2)\, P(\overline{W}_4/W_2 W_3)\, P(W_5/W_2 W_3 \overline{W}_4) + P(W_2)\, P(\overline{W}_3/W_2)\, P(W_4/W_2 \overline{W}_3)\, P(W_5/W_2 \overline{W}_3 W_4) + P(\overline{W}_2)\, P(W_3/\overline{W}_2)\, P(W_4/\overline{W}_2 W_3)\, P(W_5/\overline{W}_2 W_3 W_4) + P(W_2)\, P(W_3/W_2)\, P(W_4/W_2 W_3)\, P(W_5/W_2 W_3 W_4)$$

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

$$= \frac{8+4+4+4+16}{81} = \frac{36}{81}.$$

NOTES

EXERCISE 9.6

1. Two cards are drawn from a pack of cards in succession (with replacement). Find the probability that the first card is spade and the second is a black king.
2. A husband and a wife appear in an interview for two vacancies for the same post. The probability of husband's selection is $\frac{2}{5}$ and that of wife is $\frac{4}{5}$. What is the probability that both of them will be selected?
3. A man wants to marry a girl having qualities : white complexion—the probability of getting such a girl is one in twenty, handsome dowry — the probability of getting this is one in thirty. Find the probability of his getting married to a white complexioned girl who may also bring handsome dowry.
4. (*i*) A problem in statistics is given to three students Ram, Shyam and Radheyshyam whose chances of solving it are 0.3, 0.5 and 0.6 respectively. Find the probability that the problem will be solved.

   (*ii*) A problem in statistics is given to three students, A, B and C whose chances of solving it are 1/2, 1/3 and 1/4 respectively. Find the probability that the problem will be solved.

   (*iii*) A problem in statistics is given to four students A, B, C and D whose chances of solving it are $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{4}$ respectively. Find the probability that the problem is solved.
5. The probability of A winning a race is $\frac{1}{5}$ and the probability of B winning the race is $\frac{1}{6}$. Find the probability that none will win the race.
6. (*i*) The odds in favour of 'A' solving a problem are 7 : 6, and the odds against 'B' solving the same problem are 11 : 8. What is the probability that the problem will be solved, if both try the problem?

   (*ii*) A can solve 90% of problems given in a book and B can solve 70%. What is the probability that at least one of them will solve the problem, selected at random?
7. The odds in favour of first s peaking the truth are 3 : 2. The odds in favour of second speaking the truth are 5 : 3. In what percentage of cases are they likely to contradict each other on an identical point?
8. What is the probability of throwing 6 with a die at least once in 3 attempts?
9. A can solve 75% of problems and B can solve 70%. What is the probability that at least one of them will solve the problem, selected at random.
10. Find the probability of drawing a heart on each of the two consecutive draws of one card from a well-shuffled pack of playing cards, if the card is not replaced after the first draw.
11. Find the probability of drawing a king, a queen and a knave in that order from a pack of playing cards in three consecutive draws of one card. The first two cards drawn are replaced.
12. A bag contains 10 red and 6 black balls, 4 balls are drawn successively one by one and are not replaced. What is the probability that these are alternatively of different colours?
13. A bag contains 13 balls numbered from 1 to 13. Suppose an even number is considered as a success. Two balls are drawn one by one without replacement. Find the probability of getting one success.
14. A student is trying to seek admission in either of the two colleges. The probability that he is admitted in first college is 3/5 and that in second college is 1/3. Find the probability that he is admitted at least one of the colleges.

15. A bag contains 2 white balls and 3 black balls. Four persons, A, B, C, D in the order named each draws one ball and does not replace it. The first to draw a white ball receive ₹ 50. Determine their expectations.

[Hint. Let A, B, C, D themselves denote the probability of their winning.

$$\therefore \quad P(A) = \frac{2}{5}$$

$$P(B) = \frac{3}{5}\cdot\frac{2}{4} = \frac{3}{10}$$

$$P(C) = \frac{3}{5}\cdot\frac{2}{4}\cdot\frac{2}{3} = \frac{1}{5}$$

$$P(D) = \frac{3}{5}\cdot\frac{2}{4}\cdot\frac{1}{3}\cdot\frac{2}{2} = \frac{1}{10}$$

$\therefore$ Their respective expactations are $₹\left(50\times\frac{2}{5}\right)$, $₹\left(50\times\frac{3}{10}\right)$, $₹\left(50\times\frac{1}{5}\right)$, $₹\left(50\times\frac{1}{10}\right)$].

Answers

1. $\frac{1}{104}$ **2.** $\frac{8}{25}$ **3.** $\frac{1}{600}$

4. (*i*) 0.86 (*ii*) $\frac{3}{4}$ (*iii*) $\frac{13}{16}$

5. $\frac{4}{5}\times\frac{5}{6} = \frac{2}{3}$ **6.** (*i*) $\frac{181}{247}$ (*ii*) $\frac{97}{100}$

7. $\left[\left(\frac{3}{5}\times\frac{3}{8}\right)+\left(\frac{2}{5}\times\frac{5}{8}\right)\right]100\% = 47.5\%$ **8.** $1-\left(\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}\right) = \frac{91}{216}$

9. $\frac{37}{40}$ **10.** $\frac{13}{52}\times\frac{12}{51} = \frac{1}{17}$ **11.** 0.000455

12. $\frac{45}{364}$ **13.** $\frac{6}{13}\times\frac{7}{12}+\frac{7}{13}\times\frac{6}{12} = \frac{7}{13}$

14. $\frac{11}{15}$ **15.** ₹ 20, ₹ 15, ₹ 10, ₹ 5.

9.20. TOTAL PROBABILITY RULE

Let $E_1, E_2, \ldots, E_n$ be n mutually exclusive and exhaustive events, with non-zero probabilities, of a random experiment. If A be any arbitrary event of the sample space of the above random experiment with P(A) > 0, then

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

Proof. Let S be the sample space of the random experiment.

Since $E_1, E_2, \ldots, E_n$ are exhaustive, we have

$$S = E_1 \cup E_2 \cup \ldots \cup E_n$$

Now $$A = S \cap A = (E_1 \cup E_2 \cup \ldots \cup E_n) \cap A$$

$\Rightarrow$ $$A = (E_1 \cap A) \cup (E_2 \cap A) \cup \ldots \cup (E_n \cap A) \quad \ldots(1)$$

Since $E_1, E_2, \ldots, E_n$ are mutually exclusive, we have

$$E_i \cap E_j = \phi \text{ for } i \neq j.$$

Now $(E_i \cap A) \cap (E_j \cap A) = (E_i \cap E_j) \cap A = \phi \cap A = \phi$

NOTES

$\therefore$ $E_1 \cap A, E_2 \cap A, \ldots\ldots, E_n \cap A$ are also mutually exclusive.

By **addition theorem,** (1) implies

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

$$\Rightarrow \quad \mathbf{P(A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + \ldots\ldots P(E_n)P(A/E_n).}$$

Remark. In practical problems, it is found convenient to write as follows:

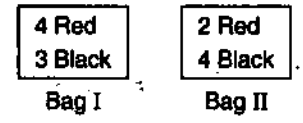
$$P(A) = P(E_1A \text{ or } E_2A \text{ or } \ldots\ldots \text{ or } E_nA) = P(E_1A) + P(E_2A) + \ldots\ldots + P(E_nA)$$

$$\therefore \quad \mathbf{P(A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + \ldots\ldots + P(E_n)P(A/E_n).}$$

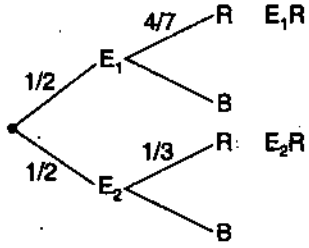
Corollary. In particular if $n = 2$, we have

$$\mathbf{P(A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2).}$$

Example 9.38. *A bag contains 4 red and 3 black balls. A second bag contains 2 red and 4 black balls. One bag is selected at random. From the selected bag, one ball is drawn. Find the probability that it is a red ball.*



Solution. Let E_1 and E_2 be the events of selecting first bag and second bag respectively.



$$\therefore \quad P(E_1) = \frac{1}{2}, \; P(E_2) = \frac{1}{2}$$

Let R be the event of drawing a red ball.

$$\therefore \quad P(R/E_1) = P(\text{Red ball is drawn from first bag}) = \frac{4}{7}$$

$$\text{Similarly,} \quad P(R/E_2) = \frac{2}{6} = \frac{1}{2}$$

Now, P(selecting a red ball)

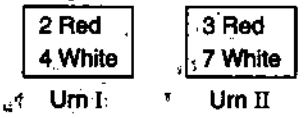
$$= P(R) = P(E_1R \text{ or } E_2R) = P(E_1R) + P(E_2R)$$

$$= P(E_1)P(R/E_1) + P(E_2)P(R/E_2)$$

$$= \frac{1}{2} \times \frac{4}{7} + \frac{1}{2} \times \frac{1}{3} = \frac{2}{7} + \frac{1}{6} = \mathbf{\frac{19}{42}}.$$

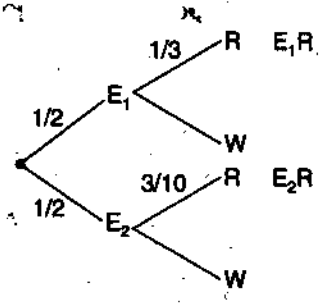
Example 9.39. *Two urns contains 2 red, 4 white and 3 red, 7 white balls. One of the urns is chosen at random and a ball is drawn from it. Find the probability that the ball drawn is (i) red (ii) white.*

Solution. Let E_1 and E_2 be the events of choosing the first urn and the second urn respectively.



$$\therefore \quad P(E_1) = \frac{1}{2}, \; P(E_2) = \frac{1}{2}$$

(*i*) Let R be the event of drawing a red ball.



$$\therefore \quad P(R/E_1) = \frac{2}{2+4} = \frac{1}{3}$$

$$P(R/E_2) = \frac{3}{3+7} = \frac{3}{10}$$

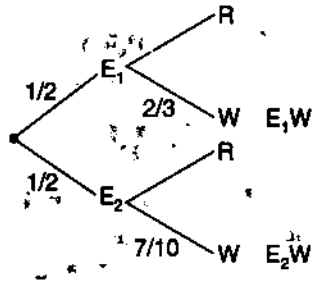
Now P(drawing a red ball) = P(R)

$$= P(E_1R \text{ or } E_2R) = P(E_1R) + P(E_2R)$$

NOTES

$$= P(E_1)P(R/E_1) + P(E_2)P(R/E_2)$$

$$= \frac{1}{2}\times\frac{1}{3} + \frac{1}{2}\times\frac{3}{10} = \frac{19}{60}.$$



(*ii*) Let W be the event of drawing a white ball.

$$\therefore \quad P(W/E_1) = \frac{4}{2+4} = \frac{2}{3}$$

$$P(W/E_2) = \frac{7}{3+7} = \frac{7}{10}$$

Now P(drawing a white ball)

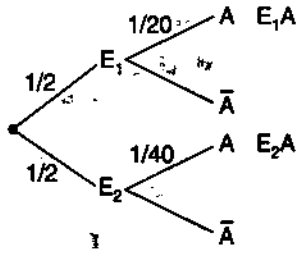
$$= P(W) = P(E_1W \text{ or } E_2W) = P(E_1W) + P(E_2W)$$

$$= P(E_1)P(W/E_1) + P(E_2)P(W/E_2)$$

$$= \frac{1}{2}\times\frac{2}{3} + \frac{1}{2}\times\frac{7}{10} = \frac{41}{60}.$$

Example 9.40. *Suppose that 5 men out of 100 men and 25 women out of 1000 women are good oretor. Assuming that there are equal number of men and women, find the probability, of choosing an oretor.*

Solution. Let E_1 and E_2 be the events of choosing a man and a woman respectively



$\therefore \quad P(E_1) = \frac{1}{2}$ and $P(E_2) = \frac{1}{2}$, because there are equal number of men and women.

Let A be the event of choosing an oretor

$\therefore \quad P(A/E_1)$ = probability that a man is an oretor

$$= \frac{5}{100} = \frac{1}{20}$$

$P(A/E_2)$ = probability that a woman is an oretor

$$= \frac{25}{1000} = \frac{1}{40}$$

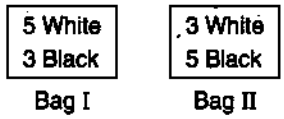
$\therefore$ P(oretor is chosen) = P(A) = P(E_1A or E_2A)

$$= P(E_1A) + P(E_2A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2)$$

$$= \frac{1}{2}\times\frac{1}{20} + \frac{1}{2}\times\frac{1}{40} = \frac{3}{80}.$$

Example 9.41. *There are two bags. The first bag contains 5 white and 3 black balls and the second bag contains 3 white and 5 black balls. Two balls are drawn at random from the first bag and are put into the second bag, without noticing their colours. Then two balls are drawn from the second bag. Find the probability that these balls are white and black.*

Solution. Let E_1, E_2 and E_3 be the events of transferring 2 white, 1 white and 1 black, 2 black balls respectively from the first bag to second bag.

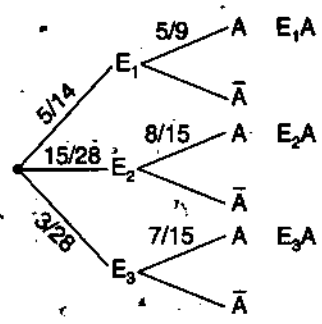


$$P(E_1) = \frac{{}^5C_2}{{}^8C_2} = \frac{10}{28} = \frac{5}{14}$$

NOTES

$$P(E_2) = \frac{{}^5C_1 \times {}^3C_1}{{}^8C_2} = \frac{5 \times 3}{28} = \frac{15}{28}$$

$$P(E_3) = \frac{{}^3C_2}{{}^8C_2} = \frac{3}{28}.$$



Let A be the event of drawing one white and one black ball from the second bag.

$$\therefore \quad P(A) = P(E_1A \text{ or } E_2A \text{ or } E_3A)$$

$$= P(E_1A) + P(E_2A) + P(E_3A)$$

$$= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + P(E_3)P(A/E_3)$$

$$= \frac{5}{14} \times \frac{{}^5C_1 \times {}^5C_1}{{}^{10}C_2} + \frac{15}{28} \times \frac{{}^4C_1 \times {}^6C_1}{{}^{10}C_2} + \frac{3}{28} \times \frac{{}^3C_1 \times {}^7C_1}{{}^{10}C_2}$$

$$= \frac{5}{14} \times \frac{5}{9} + \frac{15}{28} \times \frac{8}{15} + \frac{3}{28} \times \frac{7}{15} = \mathbf{\frac{673}{1260}}.$$

Example 9.42. *Two machines A and B produce respectively 60% and 40% of the total numbers of a items of a factory. The percentages of defective output of these machines are respectively 2% and 5%. If an item is selected at random, what is the probability that the item is (i) defective (ii) non-defective?*

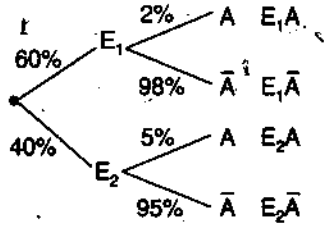
Solution. Let E_1, E_2 be the events of drawing an item produced by machine A and machine B respectively. Let A be the event of selecting a defective item.

$\therefore$ $\overline{A}$ represent the event of selecting a non-defective item. We have

$P(E_1) = 60\%$, $P(E_2) = 40\%$.

$P(A/E_1)$ = probability that a defective item is produced by A = 2%.

$P(A/E_2)$ = probability that a defective item is produced by B = 5%.



(*i*) P(selected item is defective)

$$= P(A) = P(E_1A \text{ or } E_2A) = P(E_1A) + P(E_2A)$$

$$= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) = (60\%)(2\%) + (40\%)(5\%)$$

$$= \frac{60}{100} \times \frac{2}{100} + \frac{40}{100} \times \frac{5}{100} = \frac{320}{10000} = \mathbf{0.032.}$$

(*ii*) P(selected item is non-defective)

$$= P(\overline{A}) = P(E_1\overline{A} \text{ or } E_2\overline{A}) = P(E_1\overline{A}) + P(E_2\overline{A})$$

$$= P(E_1)P(\overline{A}/E_1) + P(E_2)P(\overline{A}/E_2) = (60\%)(98\%) + (40\%)(95\%)$$

$$= \frac{60}{100} \times \frac{98}{100} + \frac{40}{100} \times \frac{95}{100} = \frac{9680}{10000} = \mathbf{0.968.}$$

EXERCISE 9.7

1. A bag contains 3 white and 2 black balls and another bag contains 2 white and 4 black balls. One bag is chosen at random. From the selected bag, one ball is drawn. Find the probability that the ball drawn is white.

NOTES

2. Find the probability of drawing a one-rupee coin from a purse with two compartments one of which contains 3 fifty-paise coins and 2 one-rupee coins and other contains 2 fifty-paise coins and 3 one-rupee coins.
3. An unbiased coin is tossed. If the result is a head, a pair of unbiased dice is rolled and the sum of the numbers obtained is noted. If the result is a tail, a card from a well shuffled pack of eleven cards numbered 2, 3, 4, ..., 12 is picked and the number on the card is noted. What is the probability that the noted number is either 7 or 8 ?
4. In a bolt factory, machines A, B and C manufacture 25%, 35% and 40% of the total bolts. Of their outputs 5%, 4% and 2% are respectively defective bolts. A bolt is drawn at random from the output. What is the probability that the bolt drawn is defective?
5. We are given three boxes as follows:
   Box I has 10 light bulbs of which 4 are defective.
   Box II has 6 light bulbs of which 1 is defective.
   Box III has 8 light bulbs of which 3 are defective.
   We select a box at random and then draw a bulb at random. What is the probability that the bulb is defective?
6. A bag contains 6 white and 7 black balls, and another bag contains 4 white and 5 black balls. A ball is taken from the first bag and without seeing its colour is put in the second bag. A ball is taken from the latter. Find the probability that the ball drawn is white.
7. Bag A contains 5 white and 6 black balls. Bag B contains 4 white and 3 black balls. A ball is transferred from bag A to the bag B and then a ball is taken out of the second bag. Find the probability of this ball being black.
8. A bag contains 3 white and 5 black balls and a second bag contains 5 white and 3 black balls. One ball is transferred from first bag to the second and then a ball is drawn from the second bag. Find the probability that the ball drawn white.
9. An urn contains 10 white and 3 black balls, while another urn contains 3 white and 5 black balls. Two balls are drawn from the first urn and put in to the second urn and then a ball is drawn from the latter. What is the probability of drawing a white ball?

Answers

1. $\frac{7}{15}$ 2. $\frac{1}{2}$ 3. $\frac{193}{792}$ 4. 0.345
5. $\frac{113}{360}$ 6. $\frac{29}{65}$ 7. $\frac{39}{88}$ 8. $\frac{43}{72}$
9. $\frac{59}{130}$

I. BAYES' THEOREM

9.21. MOTIVATION

Let there be two or more urns, each containing some white balls and red balls. Suppose an urn is chosen at random and a ball is drawn from that chosen urn. By using *addition theorem* and *multiplication theorem*, we can find the probability of drawing a white ball (or red ball) from the urn chosen.

But in case, we are given that the ball drawn is white and we are interested in finding the probability of the event that the ball was drawn from the Ist urn or IInd urn, etc., then the situation is not the same as in the previous case. Now the probability of the drawn urn will depend upon the colour of the drawn ball.

NOTES

To tackle this type of problems, Bayes' theorem is used. This theorem was enunciated by British mathematician **Thomos Bayes** in 1763.

Let $E_1, E_2, \ldots\ldots, E_n$ be n mutually exclusive and exhaustive events, with non-zero probabilities, of a random experiment. If A be any arbitrary event of the sample space of the above experiment with P(A) > 0, then

$$P(E_i/A) = \frac{P(E_i)P(A/E_i)}{\sum_{j=1}^{n} P(E_j)P(A/E_j)}, \quad 1 \le i \le n.$$

Proof. Let S be the sample space of the random experiment.

$\therefore$ $S = E_1 \cup E_2 \cup \ldots\ldots \cup E_n$ ($\because$ $E_1, E_2, \ldots\ldots, E_n$ are exhaustive)

Now $A = S \cap A = (E_1 \cup E_2 \cup \ldots\ldots \cup E_n) \cap A$

$= (E_1 \cap A) \cup (E_2 \cap A) \cup \ldots\ldots \cup (E_n \cap A)$.

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

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

or $$P(A) = \sum_{j=1}^{n} P(E_j)P(A/E_j) \qquad \ldots(1)$$

Now, $$P(E_i/A) = \frac{P(E_i \cap A)}{P(A)}, \; 1 \le i \le n$$

$\therefore$ $$P(E_i/A) = \frac{P(E_i)P(A/E_i)}{\sum_{j=1}^{n} P(E_j)P(A/E_j)}, \; 1 \le i \le n.$$ [Using (1)]

Remark 1. If $n = 2$, then

$$P(E_1/A) = \frac{P(E_1)P(A/E_1)}{P(E_1)P(A/E_1) + P(E_2)P(A/E_2)}$$

and $$P(E_2/A) = \frac{P(E_2)P(A/E_2)}{P(E_1)P(A/E_1) + P(E_2)P(A/E_2)}$$

Example 9.43. *In 1988, there will be three candidates for the position of principal – A, B and C. The chances of their selection are in the proportion 4 : 2 : 3 respectively. The probability that A, if selected, will introduce co-education in the college is 0.3. The probability of B and C doing the same are respectively 0.5 and 0.8. What is the probability that there will be co-education in the college in 1988 ? Also find the probability that the co-education in the college was introduced by the principal B.*

Solution. Let E_1, E_2, E_3 be the events of selection of A, B, C as principal respectively. Let A be the event of introduction of co-education in the college.

$\therefore$ $$P(E_1) = \frac{4}{4+2+3} = \frac{4}{9}, \; P(E_2) = \frac{2}{4+2+3} = \frac{2}{9}$$

and $$P(E_3) = \frac{3}{4+2+3} = \frac{3}{9}$$

Also, $$P(A/E_1) = \frac{3}{10}, \quad P(A/E_2) = \frac{5}{10}, \quad P(A/E_3) = \frac{8}{10}$$

Now, P(co-education is introduced in the college)

$= P(A) = P(E_1A \text{ or } E_2A \text{ or } E_3A) = P(E_1A) + P(E_2A) + P(E_3A)$

$= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + P(E_3)P(A/E_3)$

$$= \left(\frac{4}{9} \times \frac{3}{10}\right) + \left(\frac{2}{9} \times \frac{5}{10}\right) + \left(\frac{3}{9} \times \frac{8}{10}\right) = \frac{46}{90} = \frac{23}{45}.$$

By *Bayes' theorem*,

P(Co-education was introduced by the principal B)

$$= P(E_2/A) = \frac{P(E_2)P(A/E_2)}{P(A)} = \frac{\frac{2}{9} \times \frac{5}{10}}{\frac{23}{45}} = \frac{5}{23}.$$

NOTES

Example 9.44. *A manufacturing firm produces steel pipes in three plants with daily production volume of 500, 1000 and 2000 units respectively. According to past experience, it is known that the fraction of defective outputs produced by the three plants are respectively 0.005, 0.008, 0.010. If a pipe is selected from a day's total production and found to be defective, find out the probability that it came from the first plant.*

Solution. Let E_1, E_2 and E_3 be the events of drawing a pipe produced by first plant, second plant and third plant respectively. Let A be the event of drawing a defective pipe.

$$\therefore \quad P(E_1) = \frac{500}{500 + 1000 + 2000} = \frac{1}{7}$$

$$P(E_2) = \frac{1000}{500 + 1000 + 2000} = \frac{2}{7} \quad \text{and} \quad P(E_3) = \frac{2000}{500 + 1000 + 2000} = \frac{4}{7}$$

Also $P(A/E_1) = 0.005$, $P(A/E_2) = 0.008$ and $P(A/E_3) = 0.010$.

The events E_1, E_2, E_3 are mutually exclusive and exhaustive.

By *Bayes' theorem*, P(Plant I produced the defective pipe) = $P(E_1/A)$

$$= \frac{P(E_1)P(A/E_1)}{P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + P(E_3)P(A/E_3)}$$

$$= \frac{\frac{1}{7}(0.005)}{\frac{1}{7}(0.005) + \frac{2}{7}(0.008) + \frac{4}{7}(0.010)}$$

$$= \frac{0.005}{0.005 + 0.016 + 0.040)} = \frac{0.005}{0.061} = \frac{5}{61}.$$

Example 9.45. *An insurance company insured 2000 scooter drivers, 4000 car drivers and 6000 truck drivers. The probability of an accident involving a scooter driver, car driver and a truck driver is 0.01, 0.03 and 0.15 respectively. One of the insured drivers meets with an accident. What is the probability that he is a car driver?*

Solution. Let E_1, E_2, E_3 be the events of drawing scooter driver, car driver, truck driver respectively.

Total number of drivers = 2000 + 4000 + 6000 = 12000

$$\therefore \quad P(E_1) = \frac{2000}{12000} = \frac{1}{6}, \quad P(E_2) = \frac{4000}{12000} = \frac{1}{3} \quad \text{and} \quad P(E_3) = \frac{6000}{12000} = \frac{1}{2}$$

Let A be the event of getting an accident.

$\therefore$ $P(A/E_1) = 0.01$, $P(A/E_2) = 0.03$ and $P(A/E_3) = 0.15$

The events E_1, E_2 and E_3 are mutually exclusive and exhaustive.

NOTES

∴ By **Bayes' theorem**

$$P(\text{accident involved car driver}) = P(E_2/A)$$

$$= \frac{P(E_2)P(A/E_2)}{P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + P(E_3)P(A/E_3)}$$

$$= \frac{\frac{1}{3} \times 0.03}{\frac{1}{6} + 0.01 + \frac{1}{3} \times 0.03 + \frac{1}{2} \times 0.15} = \frac{0.01}{0.526} = \frac{0.06}{0.52} = \frac{3}{26}.$$

EXERCISE 9.8

1. A bag contains 4 black and 1 white balls and another bag contains 5 black and 4 white balls. One bag is chosen and a ball is drawn. If the ball drawn is black, find the probability that it is drawn from the first bag.
2. Two urns contain 4 white, 6 blue and 4 white, 5 blue balls. One of the urns is selected at random and a ball is drawn. If the ball drawn is white, find the probability that it is drawn from the second urn.
3. Assume that a factory has two machines. Past records shows that machine I produces 60% of the items of output and machine II produces 40% of the items. Further, 2% of the items produced by machine I were defective and only 1% produced by machine II were defective. If a defective item is drawn at random, what is the probability that it was produced by machine I?
4. In a factory, machines A, B and C produces 40%, 40% and 20% respectively. Of the total of their output 1%, 1% and 3% are defective. An item is drawn at random from the total production and found to be defective. Find the probability that this item is produced by the machine C.
5. A manufacturing firm produces pipes in two plants with daily production volume of 5000 and 7000 units respectively. According to past experience, it is known that the fraction of defective outputs produced by the plants are 0.01 and 0.02 respectively. If a pipe is selected at random from a day's total production and found to be defective, find out the probability that it came from the second plant.

Answers

1. 36/61 2. 10/19 3. 3/4 4. 3/7
5. 14/19

9.22. CRITICISM OF CLASSICAL APPROACH OF PROBABILITY

Though the classical approach of measuring probability seems to be quite simple and straight forward, but this approach is subjected to certain points of criticism.

In defining probability of an event, we assume that all the possible outcomes are equally likely. This means that all the possible outcomes of an experiment have equal chances of being occurred. In other words, the probability of occurrence of each outcome is equal. Thus, in classical approach, we define probability of an event in terms of probabilities of various outcomes of the experiment. Thus, this definition of probability is circular in nature.

The classical approach is based on abstract reasoning and is suitable under ideal conditions. For example, we say that in the experiment of throwing a die, the probability of getting 2 is 1/6. But this will be true if the die is unbiased *i.e.* it is perfect. In practice, perfection is not achieved. Thus, this approach is not realistic in nature.

NOTES

9.23. EMPIRICAL APPROACH OF PROBABILITY

This approach is based upon repetitive experiments under uniform conditions. Suppose a coin is perfectly balanced and we toss it 100 times. In 100 tosses, we may get head 56 times. Again if we toss this coin 1000 times, we may get head 519 times. Again if we toss this coin 10,000 times, we may get head 5085 times. In these experiments, we see that the ratio 56/100, 519/1000, 5085/10000 is tending toward 1/2, which should be the probability of getting head in any toss of the coin. In empirical approach, the probability of an events is defined in terms of a ratio of the type explained above.

If an experiment is repeated n times under uniform conditions and an event E occurs 'm' times, then the probability of the event E is defined as

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

In the definition, '$\lim_{n\to\infty}$' emphasizes the fact that n must be very large. In the real mathematical sense, we cannot measure $\lim_{n\to\infty} \frac{m}{n}$, because we cannot repeat any experiment infinitely many times. Thus, in this approach, we content ourselves by assuming that n takes large and practically possible values. If we toss a perfect coin two times, it is not expected that we shall definitely get one head and one tail. But if the same coin is tossed 5000 times, we may get about 2500 heads. Thus, the probability defined in this approach is a long run concept, because to find the probability of an event, we have to repeat out experiment a large number of times. In any experiment, we shall have $m \le n$.

$$\therefore \quad P(E) = \lim_{n\to\infty} \frac{m}{n} \text{ implies } \quad 0 \le P(E) \le 1.$$

The empirical approach of probability is based on experience.

Limitations. The computation of empirical probability requires repetition of experiment, a very large number of times. This restrict the suitability of this approach. In many cases, the experiment may not be repeated large number of times. If E be the event that a particular student secures 70% marks or more in all the examinations given to him in a particular year, then this experiment cannot be repeated many number of times. This approach is also not applicable to experiments which are not expected to occur frequently in future.

NOTES

9.24. SUBJECTIVE APPROACH OF PROBABILITY

The classical and empirical approaches of probability are *objective* in nature. In the subjective approach, the probability of an event is considered as a measure of one's confidence in the occurrence of that particular event. The probability of the event that the student 'A' will pass the examination cannot be calculated by any of the above discussed objective approaches. The events of his passing and failing are not equally likely cases. Had these cases been equally likely, we could have used classical approach and said that the probability is $\frac{1}{2}$. In this case, the experiment is such that it cannot be repeated under uniform conditions. Thus, the empirical approach also fails to comment upon the probability of this event. In such cases, the subjective approach is found useful. In subjective approach, the probability of an event represent the degree of faith which a rational person reposes in the occurrence of that certain event. The degree of faith will depend upon his judgement, personal outlook, etc. In this approach, the probability of a event, differ from person to person and that is why it is called subjective approach. In this approach, the probability of an event also suffer from personal bias of its estimator.

9.25. SUMMARY

- When we perform experiments in science and engineering, repeatedly under very nearly identical conditions, we get almost the same result. Such experiments are called **deterministic experiments**.

  There also exist experiments in which the results may not be essentially the same even if the experiment is performed under very nearly identical conditions. Such experiments are called **random experiments**.
- The **sample space** of a random experiment is defined as the set of all possible outcomes of the experiment. The possible outcomes are called **sample points.**
- A **Tree diagram** is a device used to enumerate all the logical possibilities of a sequence of steps where each step can occur in a finite number of ways.
- An **event** is defined as a subset of the sample space. An event is called an **elementary (or simple) event** if it contains only one sample point. In the experiment of rolling a die, the event A of getting '3' is a simple event. We write A = {3}. An event is called an **impossible event** if it can never occur.

9.26. REVIEW EXERCISES

1. Explain the fundamental concepts of 'Probability'.
2. Define 'probability'.
3. Write the fundamental concepts of probability calculation.
4. Define 'probability' and explain its importance in Statistics.
5. Explain the term 'Mutually Exclusive Events' by taking some examples.
6. What is conditional probability? Explain with the help of an example.
7. Define probability and explain the Addition law of probability giving suitable examples.

NOTES

8. Explain what do you understand by the term 'probability'. State and prove the addition and multiplication theorems of probability.
9. Explain short notes on any two:
   (*i*) Dependent and independent events
   (*ii*) Mutually exclusive and equally likely events
   (*iii*) Simple and compound events.
10. Explain the Multiplication Theorem of Probability with suitable example.
11. Explain Bayes' theorem with the help of an example.
12. Define probability in different ways. Giving their merits and demerits by examples. State which is the best.
13. Discuss in detail the Classical and Empirical approaches to probability.
14. Explain the various approaches to probability.

NOTES

10. PROBABILITY DISTRIBUTIONS (Binomial, Poisson, Normal Distributions)

STRUCTURE

NOTES

10.1. INTRODUCTION

We know that a real valued function defined on the sample space of a random experiment is called a *random variable*. A random variable is either discrete or continuous.

Let x be a discrete random variable assuming values $x_1, x_2, x_3, \ldots, x_n$ corresponding to the various outcomes of a random experiment. If the probability of occurrence of $x = x_i$ is $P(x_i) = p_i, 1 \le i \le n$ such that $p_1 + p_2 + p_3 + \ldots + p_n = 1$, then the function, $P(x_i) = p_i, 1 \le i \le n$ is called the *probability function* of the random variable x and the set $\{P(x_1), P(x_2), P(x_3), \ldots, P(x_n)\}$ is called the *probability distribution* of x.

10.2. EMPIRICAL DISTRIBUTION

Let x be a discrete random variable assuming values $x_1, x_2, \ldots, x_n$ corresponding to various outcomes of a random experiment. Let this random experiment be repeated N times. Let the random variable x take values $x_1, x_2, \ldots, x_n$ with respective frequencies $f_1, f_2, \ldots, f_n$ where $f_1 + f_2 + \ldots + f_n = N$.

The distribution

x	x_1	x_2	....	x_n
f	f_1	f_2	....	f_n

is called an **empirical distribution.**

Illustration. Let the random experiment be of tossing of two coins.

$\therefore$ S = {HH, HT, TH, TT}.

Let x be random variable "square of number of tails," then x takes the values $0^2 = 0$, $1^2 = 1$, $2^2 = 4$. Let this random experiment be repeated 100 times and let the observed frequencies be as follows:

HH	HT	TH	TT
↓	↓	↓	↓
24	27	23	26

$\therefore$ The empirical distribution corresponding to above experiment is

x	0 (HH)	1 (HT, TH)	4 (TT)
f	24	50 (= 27 + 23)	26

Now we shall consider three very important types of probability distributions.

I. BINOMIAL DISTRIBUTION

10.3. INTRODUCTION

The binomial distribution is a particular type of probability distribution. This was discovered by **James Bernoulli (1654—1705)** in the year 1700. This distribution mainly deals with attributes. An attribute is either present or absent with respect to

NOTES

elements of a population. For example, if a coin is tossed, we get either *head* or *tail*. The workers of a factory may be classified as *skilled* and *unskilled*. An item of a population of articles produced in a firm may be either defective or non-defective.

10.4. CONDITIONS

The following conditions are essential for the applicability of binomial distribution:

(*i*) **The random experiment is performed for a finite and fixed number of trials.** If in an experiment, a coin is tossed repeatedly or a ball is drawn from an urn repeatedly, then each toss or draw is called a **trial.** For example, if a coin is tossed 6 times, then this experiment has 6 trials. The number of trials is an experiment is generally denoted by '*n*'.

(*ii*) **The trials are independent.** By this we mean that the result of a particular trial is not going to effect the result of any other trial. For example, if a coin is tossed or a die is thrown, the trials would be independent. If from a pack of playing cards, some draws of one card are made without replacing the cards, then the trials would not be independent. But if the card drawn is replaced before the next draw, then the trials would be independent.

(*iii*) **Each trial must result in either "success" or "failure".** In other words, in every trial, there should be only two possible outcomes *i.e., success* or *failure*. For example, if a coin is tossed, then either *head* or *tail* is observed. Similarly, if an item is examined, it is either *defective* or *non-defective*.

(*iv*) **The probability of success in each trial is same.** In other words, this condition requires that the probability of *success* should not change in different trials. For example, if a sample of two items is drawn, then the probability of exactly one being defective will be constant in different trials provided the items are replaced before the next draw.

10.5. BINOMIAL VARIABLE

A random variable which counts the number of successes in a random experiment with trials satisfying above four conditions is called a **Binomial variable.**

For example, if a coin is tossed 5 times and the event of getting head is *success*, then the possible values of the binomial variable are 0, 1, 2, 3, 4, 5. This is so, because, the minimum number of successes is 0 and maximum number is 5.

10.6. BINOMIAL PROBABILITY FUNCTION

When a fair coin is tossed, we have only two possibilities: head and tail. Let us call the occurrence of head as 'success'. Therefore, the occurrence of tail would be a 'failure'. Let this coin be tossed 5 times. Suppose we are interested in finding the probability of getting 4 heads and 1 tail *i.e.*, of getting 4 successes. If S and F denote 'success' and 'failure' in a trial respectively, then there are ${}^5C_4 = 5$ ways of having 4 successes.

These are: SSSSF, SSSFS, SSFSS, SFSSS, FSSSS.

NOTES

The probability of getting 4 successes in each case is $\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)$, because the trials are independent.

∴ By using *addition theorem,* the required probability of having 4 successes is ${}^5C_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)$, which is equal to $\frac{5}{32}$. Now we shall generalise this method of finding the probabilities for different values of *binomial variables.*

Let a random experiment satisfying the conditions of **binomial distribution** be performed. Let the number of trials in the experiment be n. Let p denote the probability of *success* in any trial.

∴ Probability of failure, $q = 1 - p$

Let x denote the binomial variable corresponding to this experiment.

∴ The possible values of x are 0, 1, 2, ..., n.

If there are r successes in n trials, then there would be $n - r$ failures. One of the ways in which r successes may occur is

$$\underbrace{SS \ldots\ldots SS}_{r \text{ times}} \quad \underbrace{FF \ldots\ldots F}_{n-r \text{ times}}$$

where S and F denote success and failure in trials.

Now, P(SS ..... SFF ..... F) = P(S)P(S) ..... P(S)P(F)P(F) ..... P(F)

(∵ the trials are independent)

$$= p.p \ldots\ldots p.q.q \ldots\ldots q = p^r q^{n-r}.$$

We know that ${}^nC_r$ is the number of combinations of n things taking r at a time. Therefore, the number of ways in which r successes can occur in n trials is equal to the number of ways of choosing r trials (for successes) out of total n trials *i.e.*, it is ${}^nC_r$. Therefore, there are ${}^nC_r$ ways in which we get r successes and $n - r$ failures and the probability of occurrence of each of these ways is $p^r q^{n-r}$. Hence the probability of r successes in n trials in any order is

$$P(x = r) = p^r q^{n-r} + p^r q^{n-r} + \ldots\ldots {}^nC_r \text{ terms} \qquad \text{(By addition theorem)}$$

or $$\mathbf{P(x = r) = {}^nC_r p^r q^{n-r},\ 0 \le r \le n.}$$

This is called the **binomial probability function.** The corresponding **binomial distribution** is

x	0	1	2	.....	n
$P(x)$	${}^nC_0 p^0 q^n$	${}^nC_1 p^1 q^{n-1}$	${}^nC_2 p^2 q^{n-2}$	.....	${}^nC_n p^n q^0$

The probabilities of 0 success, 1 success, 2 successes, ..., n successes are respectively the 1st, 2nd, 3rd, ..., $(n + 1)$th terms in the binomial expansion of $(q + p)^n$. This is why, it is called **binomial distribution.**

10.7. BINOMIAL FREQUENCY DISTRIBUTION

If a random experiment, satisfying the requirements of binomial distribution, is repeated N times, then the expected frequency of getting r $(0 \le r \le n)$ successes is given by

$$\mathbf{N.P(x = r) = N.{}^nC_r p^r q^{n-r},\ 0 \le r \le n.}$$

NOTES

The *frequencies* of getting 0 success, 1 success, 2 successes, ..., n successes are respectively the 1st, 2nd, 3rd, ..., $(n+1)$th terms in the expansion of $N(q+p)^n$.

Example 10.1. *An unbiased coin is tossed six times. Find the probability of obtaining:*

(i) no head *(ii) all heads*

(iii) at least one head i.e., one or more heads

(iv) exactly 4 heads *(v) less than 3 heads*

(vi) more than 4 heads *(vii) more than 4 and less than 6 heads*

(viii) more than 6 heads.

Solution. Let p be the probability of success *i.e.*, of getting head in the toss of the coin.

$$\therefore \qquad n = 6, p = \frac{1}{2} \quad \text{and} \quad q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}.$$

Let x be the binomial variable **'no. of successes'.**

By **Binomial distribution**, $P(x = r) = {}^nC_r p^r q^{n-r}, \quad 0 \le r \le n.$

$$\therefore \qquad P(x = r) = {}^6C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{6-r} = {}^6C_r \left(\frac{1}{2}\right)^6 = {}^6C_r \frac{1}{64}, \quad 0 \le r \le n.$$

(i) P(no head) $= P(x = 0) = {}^6C_0 \frac{1}{64} = \mathbf{\frac{1}{64}}$

(ii) P(all heads) $= P(x = 6) = {}^6C_6 \frac{1}{64} = \mathbf{\frac{1}{64}}$

(iii) P(at least one head) $= 1 - P(\text{no head}) = 1 - \frac{1}{64} = \mathbf{\frac{63}{64}}$ [Using part *(i)*]

(iv) P(exactly 4 heads) $= P(x = 4) = {}^6C_4 \frac{1}{64} = \mathbf{\frac{15}{64}}$

(v) P(less than 3 heads) $= P(x < 3) = P(x = 0 \text{ or } 1 \text{ or } 2)$

$$= P(x = 0) + P(x = 1) + P(x = 2)$$

$$= {}^6C_0 \frac{1}{64} + {}^6C_1 \frac{1}{64} + {}^6C_2 \frac{1}{64}$$

$$= (1 + 6 + 15) \frac{1}{64} = \frac{22}{64} = \mathbf{\frac{11}{32}}$$

(vi) P(more than 4 heads)

$$= P(x > 4) = P(x = 5 \text{ or } 6) = P(x = 5) + P(x = 6)$$

$$= {}^6C_5 \frac{1}{64} + {}^6C_6 \frac{1}{64} = (6 + 1) \frac{1}{64} = \mathbf{\frac{7}{64}}$$

(vii) P(more than 4 heads and less than 6 heads)

$$= P(4 < x < 6) = P(x = 5) = {}^6C_5 \frac{1}{64} = \frac{6}{64} = \mathbf{\frac{3}{32}}.$$

(viii) P(more than 6 heads) $= P(x > 6) = \mathbf{0}.$ (∵ The event is impossible.)

Example 10.2. *A die is thrown 4 times. Getting a number greater than 2 is a success. Find the probability of getting (i) exactly 1 success (ii) less than 3 successes (iii) more than 3 successes.*

Solution. Let p be the probability of success *i.e.*, of getting number greater than 2 in the throw of a die.

NOTES

$\therefore \quad n = 4, p = \frac{4}{6} = \frac{2}{3}$ and $q = 1 - p = 1 - \frac{2}{3} = \frac{1}{3}$.

Let x be the binomial variable "*no. of successes*".

By **Binomial distribution**, $P(x = r) = {}^nC_r\, p^r q^{n-r}, 0 \le r \le n$.

$\therefore \quad P(x = r) = {}^4C_r \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{4-r}, 0 \le r \le 4.$

(*i*) P(exactly 1 success) = P(x = 1)

$$= {}^4C_1 \left(\frac{2}{3}\right)^1 \left(\frac{1}{3}\right)^{4-1} = 4 \times \frac{2}{3} \times \frac{1}{27} = \frac{8}{81}.$$

(*ii*) P(less than 3 successes) = P(x < 3)

= P(x = 0 or x = 1 or x = 2)

= P(x = 0) + P(x = 1) + P(x = 2)

(Addition theorem for m.e. events)

$$= {}^4C_0 \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^{4-0} + {}^4C_1 \left(\frac{2}{3}\right)^1 \left(\frac{1}{3}\right)^{4-1} + {}^4C_2 \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^{4-2}$$

$$= \left(1 \times 1 \times \frac{1}{81}\right) + \left(4 \times \frac{2}{3} \times \frac{1}{27}\right) + \left(6 \times \frac{4}{9} \times \frac{1}{9}\right)$$

$$= \frac{1 + 8 + 24}{81} = \frac{33}{81}.$$

(*iii*) P(more than 3 successes) = P(x > 3)

$$= P(x = 4) = {}^4C_4 \left(\tfrac{2}{3}\right)^4 \left(\tfrac{1}{3}\right)^0 = 1 \times \frac{16}{81} \times 1 = \frac{16}{81}.$$

Example 10.3. *There are 20% chances for a worker of an industry to suffer from an occupational disease. 50 workers were selected at random and examined for the occupational disease. Find the probability that (i) only one worker is found suffering from the disease; (ii) more than 3 are suffering from the disease; (iii) none is suffering from the disease.*

Solution. Let p be the probability of success *i.e.*, a worker is suffering from disease.

$\therefore \quad n = 50, p = \frac{20}{100} = \frac{1}{5}$ and $q = 1 - p = 1 - \frac{1}{5} = \frac{1}{4}$.

Let x be the binomial variable, "*no. of successes*".

By **Binomial distribution,** $P(x = r) = {}^nC_r\, p^r q^{n-r}, 0 \le r \le n$.

$\therefore \quad P(x = r) = {}^{50}C_r \left(\frac{1}{5}\right)^r \left(\frac{4}{5}\right)^{50-r}, 0 \le r \le 50.$

(*i*) P(only one is suffering) = P(x = 1)

$$= {}^{50}C_1 \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^{49} = 50 \times \frac{1}{5} \times \left(\frac{4}{5}\right)^{49} = 10 \left(\frac{4}{5}\right)^{49}.$$

(*ii*) P(more than 3 are suffering) = P(x > 3) = 1 − P($x \le 3$)

= 1 − P(x = 0 or x = 1 or x = 2 or x = 3)

= 1 − {P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)}

NOTES

$$= 1 - \left\{ {}^{50}C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{50} + {}^{50}C_1 \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^{49} + {}^{50}C_2 \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^{48} + {}^{50}C_3 \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^{47} \right\}$$

$$= 1 - \left\{ 1 \times 1 \times \left(\frac{4}{5}\right)^{50} + 50 \times \frac{1}{5} \times \left(\frac{4}{5}\right)^{49} + \left(\frac{50 \times 49}{1 \times 2}\right) \times \left(\frac{1}{5}\right)^2 \times \left(\frac{4}{5}\right)^{48} + \left(\frac{50 \times 49 \times 48}{1 \times 2 \times 3}\right) \times \left(\frac{1}{5}\right)^3 \times \left(\frac{4}{5}\right)^{47} \right\}$$

$$= 1 - \frac{4^{47}}{5^{50}} \{64 + (50 \times 16) + (1225 \times 4) + (19600 \times 1)\} = 1 - \left(\frac{4^{47}}{5^{50}} \times 25364\right).$$

(*iii*) P(none is suffering) = P($x = 0$)

$$= {}^{50}C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{50} = 1 \times 1 \times \left(\frac{4}{5}\right)^{50} = \left(\frac{4}{5}\right)^{50}.$$

Example 10.4. *There are 64 beds in a garden and 3 seeds of a particular type of flower are sown in each bed. The probability of a flower being white is 1/4. Find the number of beds with 3, 2, 1 and 0 white flowers.*

Solution. Let p be the probability of success *i.e.,* the flower is white.

$$\therefore \qquad n = 3,\ N = 64,\ p = \frac{1}{4},\ q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}.$$

Let x be the binomial variable, '*no. of successes*'.

By **Binomial distribution,** $P(x = r) = {}^nC_r p^r q^{n-r}$; $0 \le r \le n$.

$$\therefore \qquad P(x = r) = {}^3C_r \left(\frac{1}{4}\right)^r \left(\frac{3}{4}\right)^{3-r},\ 0 \le r \le 3.$$

$$\therefore \text{ P(3 white flowers)} = P(x = 3) = {}^3C_3 \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^0 = \frac{1}{64}$$

$$\text{P(2 white flowers)} = P(x = 2) = {}^3C_2 \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^1 = \frac{9}{64}$$

$$\text{P(1 white flower)} = P(x = 1) = {}^3C_1 \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^2 = \frac{27}{64}$$

$$\text{P(no white flower)} = P(x = 0) = {}^3C_0 \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^3 = \frac{27}{64}.$$

$\therefore$ Number of beds with 3 white flowers

$$= 64 \times P(x = 3) = 64 \times \frac{1}{64} = 1$$

Number of beds with 2 white flowers

$$= 64 \times P(x = 2) = 64 \times \frac{9}{64} = 9$$

Number of beds with 1 white flower

$$= 64 \times P(x = 1) = 64 \times \frac{27}{64} = 27$$

Number of beds with no white flower

$$= 64 \times P(x = 0) = 64 \times \frac{27}{64} = 27.$$

Example 10.5. *In an experiment, a fair die is thrown 6 times. The event of occurring number greater than 4 is the 'success' of the experiment. Find the Binomial distribution of the experiment. If the same experiment is repeated 3645 times, find also the Binomial frequency distribution.*

Solution. Let p be the probability of success *i.e.*, of getting number greater than 4

$$\therefore \quad n = 6, p = \frac{2}{6} = \frac{1}{3} \text{ and } q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}.$$

Let x be the binomial variable, "*no. of successes*".

By **Binomial distribution**, $P(x = r) = {}^nC_r p^r q^{n-r}$, $0 \le r \le n$.

$$P(x = r) = {}^6C_r \left(\frac{1}{3}\right)^r \left(\frac{2}{3}\right)^{6-r}, \quad 0 \le r \le 6.$$

Now
$$P(x = 0) = {}^6C_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^6 = 1 \times 1 \times \frac{64}{729} = \frac{64}{729}$$

$$P(x = 1) = {}^6C_1 \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^5 = 6 \times \frac{1}{3} \times \frac{32}{243} = \frac{192}{729}$$

$$P(x = 2) = {}^6C_2 \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^4 = 15 \times \frac{1}{9} \times \frac{16}{81} = \frac{240}{729}$$

$$P(x = 3) = {}^6C_3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^3 = 20 \times \frac{1}{27} \times \frac{8}{27} = \frac{160}{729}$$

$$P(x = 4) = {}^6C_4 \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^2 = 15 \times \frac{1}{81} \times \frac{4}{9} = \frac{60}{729}$$

$$P(x = 5) = {}^6C_5 \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right)^1 = 6 \times \frac{1}{243} \times \frac{2}{3} = \frac{12}{729}$$

$$P(x = 6) = {}^6C_6 \left(\frac{1}{3}\right)^6 \left(\frac{2}{3}\right)^0 = 1 \times \frac{1}{729} \times 1 = \frac{1}{729}.$$

$\therefore$ The required distributions are as follows:

Binomial Distribution		*Binomial Frequency Distribution*	
x	$P(x)$	x	$N.P(x) = 3645 \times P(x)$
0	$\frac{64}{729}$	0	$3645 \times \frac{64}{729} = 320$
1	$\frac{192}{729}$	1	$3645 \times \frac{192}{729} = 960$
2	$\frac{240}{729}$	2	$3645 \times \frac{240}{729} = 1200$
3	$\frac{160}{729}$	3	$3645 \times \frac{160}{729} = 800$
4	$\frac{60}{729}$	4	$3645 \times \frac{60}{729} = 300$
5	$\frac{12}{729}$	5	$3645 \times \frac{12}{729} = 60$
6	$\frac{1}{729}$	6	$3645 \times \frac{1}{729} = 5$

NOTES

EXERCISE 10.1

1. If a coin is tossed six times, what is the probability of obtaining four or more heads?
2. An unbiased coin is tossed 8 times. Find the probability of obtaining (*i*) exactly 2 heads (*ii*) more than 2 heads (*iii*) all heads.
3. The incidence of occupational disease in a factory is such that the workers have a 25% chances of suffering from it. What is the probability that out of 6 workmen, 4 or more contact the disease?
4. Out of 800 families with 4 children each, what percentage would be expected to have (*a*) 2 boys and 2 girls (*b*) at least one boy (*c*) no girl (*d*) at most 2 girls? Assume equal probabilities for boys and girls.
5. In a certain town, 20% of population is literate, and assume that 200 investigators takes a sample of 10 individuals to see whether they are literate. How many investigators would you expect to report that 3 persons or less are literate in their samples?
6. Eight coins are tossed at a time, 256 times. Find the expected frequencies of 0; 1, 2, 3 successes (getting head).
7. During war, 1 ship out of 9 was sunk on an average in making a certain voyage. What was the probability that exactly 3 out of a convoy of 6 ships would arrive safely?
8. The probability of a man hitting a target is $\frac{1}{3}$. How many least number of times, must he fire so that the probability of hitting the target at least once is more than 90%?
9. An unbiased coin is tossed 10 times. Find the probability of getting exactly 5 heads?
10. The probability that a student will be graduate is 0.4. Detremine the probability that out of 5 students (*i*) none (*ii*) one (*iii*) at least one (*iv*) all will be graduate.

Answers

1. $\frac{11}{32}$
2. $\frac{7}{64}, \frac{219}{256}, \frac{1}{256}$
3. $\frac{77}{2048}$
4. 37.5%, 93.75%, 6.25%, 68.75%
5. 176
6. 1, 8, 28, 56
7. ${}^6C_3\left(\frac{8}{9}\right)^3\left(\frac{1}{9}\right)^3 = 0.0193$
8. $1-\left(\frac{2}{3}\right)^n > \frac{9}{10} \Rightarrow n = 6$
9. 0.2461
10. (*i*) 0.07776 (*ii*) 0.2592 (*iii*) 0.92224 (*iv*) 0.01024

II. PROPERTIES OF BINOMIAL DISTRIBUTION

10.8. THE SHAPE OF B.D.

The shape of the binomial distribution depends upon the probability of success (p) and the number of trials in the experiment. If $p = q = \frac{1}{2}$, then the distribution will be symmetrical for every value of n. If $p \neq q$, then the distribution would be asymmetrical *i.e.*, skewed. The magnitude of skewness varies as the difference between p and q. We illustrate this by taking $p = 0.1$, $p = 0.5$, $p = 0.9$ and assuming that there are 5 trials in the experiment.

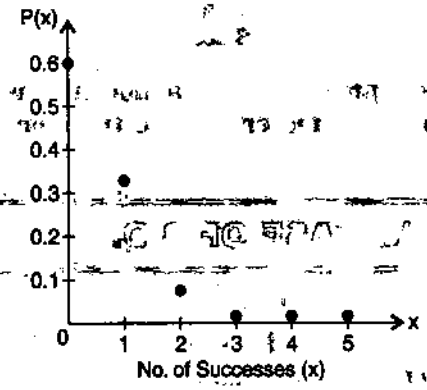
NOTES

Case I. p = 0.1, n = 5

We have $P(x = r) = {}^5C_r \left(\frac{1}{10}\right)^r \left(\frac{9}{10}\right)^{5-r}$, $0 \le r \le 5$.

The B.D. is

x	0	1	2	3	4	5
$P(x)$	0.59049	0.32805	0.07290	0.00810	0.00045	0.00001

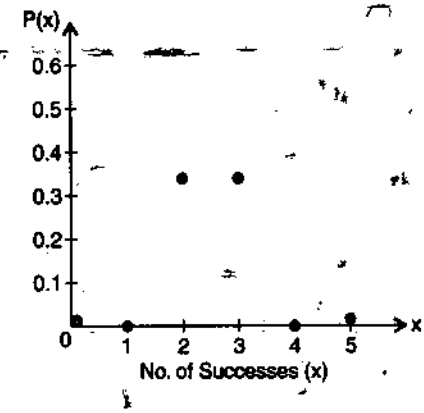


Case II. p = 0.5, n = 5

We have $P(x = r) = {}^5C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{5-r}$, $0 \le r \le 5$.

The B.D is

x	0	1	2	3	4	5
$P(x)$	0.03125	0.15625	0.31250	0.31250	0.15625	0.03125



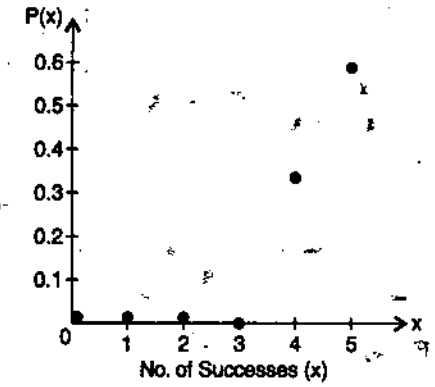
Case III. p = 0.9, n = 5

We have $P(x = r) = {}^5C_r \left(\frac{9}{10}\right)^r \left(\frac{1}{10}\right)^{5-r}$, $0 \le r \le 5$.

The B.D. is

x	0	1	2	3	4	5
$P(x)$	0.00001	0.00045	0.00810	0.07290	0.32805	0.59049

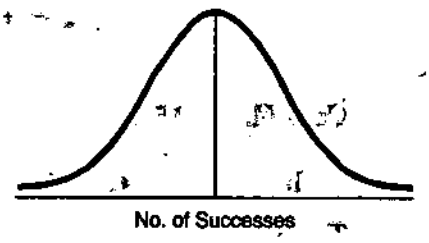
NOTES



Thus, we see that the probabilities in a binomial distribution depends upon n and p. These are called the **parameter** of the distribution.

10.9. THE LIMITING CASE OF B.D.

As number of trials (n) in the binomial distribution increases, the number of successes also increases. If neither p nor q is very small, then as n approaches infinity, the skewness in the distribution disappears and it becomes continuous. We shall see that such a continuous, bell shaped distribution is called a *normal distribution*. Thus, the normal distribution is a limiting case of binomial distribution as n approaches infinity.



10.10. MEAN OF B.D.

Let x be a binomial random variable and

$$P(x = r) = {}^nC_r p^r q^{n-r},\ 0 \le r \le n.$$

The *mean* of x is the average number of successes.

$$\therefore \text{ Mean, } \mu = \sum_{r=0}^{n} r.P(x = r) = \sum_{r=0}^{n} r.{}^nC_r p^r q^{n-r}$$

$$= 0.{}^nC_0 p^0 q^n + 1.{}^nC_1 p^1 q^{n-1} + 2.{}^nC_2 p^2 q^{n-2} + \ldots\ldots + n.{}^nC_n p^n.q^0$$

$$= 0 + n.pq^{n-1} + 2.\frac{n(n-1)}{1.2} p^2 q^{n-2} + \ldots\ldots + n.1.p^n$$

$$= np\left\{q^{n-1} + \frac{n-1}{1} pq^{n-2} + \ldots + p^{n-1}\right\}$$

$$= np\{{}^{n-1}C_0 p^0 q^{n-1} + {}^{n-1}C_1 p^1 q^{n-2} + \ldots + {}^{n-1}C_{n-1} p^{n-1} q^0\}$$

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

$\therefore$ Mean of x = **np.**

NOTES

10.11. VARIANCE AND S.D. OF B.D.

Let x be a binomial random variable and

$$P(x = r) = {}^nC_r p^r q^{n-r},\ 0 \le r \le n.$$

The variance and standard deviation of x measures the dispersion of the binomial distribution and are given by

$$\text{Variance} = \sum_{r=0}^{n} r^2 . P(x = r) - \mu^2$$

and

$$\text{S.D} = \sqrt{\sum_{r=0}^{n} r^2 . P(x = r) - \mu^2}$$

Now $$\sum_{r=0}^{n} r^2 . P(x = r) = \sum_{r=0}^{n} r^2 . {}^nC_r p^r q^{n-r}$$

$$= 0 . {}^nC_0 p^0 q^n + 1^2 . {}^nC_1 p^1 q^{n-1} + 2^2 . {}^nC_2 p^2 q^{n-2} + 3^2 . {}^nC_3 p^n q^{n-3} + \ldots\ldots + n^2 . {}^nC_n p^n q^0$$

$$= 0 + 1 . \frac{n}{1} pq^{n-1} + 2^2 . \frac{n(n-1)}{1.2} p^2 q^{n-2} + \frac{3^2 . n(n-1)(n-2)}{1.2.3} p^3 q^{n-3} + \ldots\ldots + n^2 . 1 . p^n . 1$$

$$= np\left\{q^{n-1} + \frac{2(n-1)}{1} pq^{n-2} + \frac{3(n-1)(n-2)}{1 \times 2} p^2 q^{n-3} + \ldots\ldots + np^{n-1}\right\}$$

$$= np\left\{\left(q^{n-1} + \frac{n-1}{1} pq^{n-2} + \frac{(n-1)(n-2)}{1 \times 2} p^2 q^{n-3} + \ldots\ldots + p^{n-1}\right)\right.$$

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

$$= np\{(q + p)^{n-1} + (n-1)\, p(q^{n-2} + (n-2)\, pq^{n-3} + \ldots\ldots + p^{n-2})\}$$

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

$$= np\{1 + np - p\} = np + n^2p^2 - np^2.$$

$$\therefore \quad \text{Variance} = \sum_{r=0}^{n} r^2 . P(x = r) - \mu^2 = (np + n^2p^2 - np^2) - (np)^2 = np - np^2$$

$$= np(1 - p) = \mathbf{npq}.$$

Also, $\quad \text{S.D.} = \sqrt{\textbf{variance}} = \sqrt{\textbf{npq}}$.

10.12. γ_1 AND γ_2 OF B.D.

The values of γ_1 and γ_2 for the binomial probability function

$$P(x = r) = {}^nC_r p^r q^{n-r},\ 0 \le r \le n$$

are given by

$$\gamma_1 = \frac{\mathbf{1 - 2p}}{\sqrt{\mathbf{npq}}} \quad \text{and} \quad \gamma_2 = \frac{\mathbf{1 - 6pq}}{\mathbf{npq}}.$$

NOTES

10.13. RECURRENCE FORMULA FOR B.D.

Let x be a binomial random variable and

$$P(x = r) = {}^nC_r p^r q^{n-r}\ 0 \le r \le n.$$

For $0 \le k < n$, $P(k) = {}^nC_k p^k q^{n-k}$ and $P(k+1) = {}^nC_{k+1} p^{k+1} q^{n-(k+1)}$

Dividing, we get $$\frac{P(k+1)}{P(k)} = \frac{{}^nC_{k+1} p^{k+1} q^{n-k-1}}{{}^nC_k p^k q^{n-k}}$$

$$= \frac{n!}{(k+1)!(n-(k+1))!} \cdot \frac{k!(n-k)!}{n!} \cdot \frac{p}{q} = \frac{n-k}{k+1} \frac{p}{q}.$$

$$\therefore \quad \mathbf{P(k+1) = \frac{n-k}{k+1} \cdot \frac{p}{q} P(k) \text{ for } 0 \le k < n.}$$

This the required **recurrence formula.**

Example 10.6. *The mean and S.D. of a binomial distribution are 20 and 4 respectively. Calculate n, p and q.*

Solution. Let the binomial distribution be

$$P(x = r) = {}^nC_r p^r q^{n-r},\ 0 \le r \le n.$$

$\therefore$ Mean $= np$ and S.D. $= \sqrt{npq}$.

We are given mean $= 20$, S.D. $= 4$.

$\therefore \quad np = 20,\ \sqrt{npq} = 4$

$\Rightarrow \quad \sqrt{20q} = 4 \quad \Rightarrow \quad 20q = 16 \quad \Rightarrow \quad q = \frac{4}{5}$

$\therefore \quad p = 1 - q = 1 - \frac{4}{5} = \frac{1}{5}$.

$np = 20$ implies $n \times \frac{1}{5} = 20$ *i.e.*, $n = 100$

$\therefore \quad n = 100, p = \frac{1}{5}, q = \frac{4}{5}$.

Example 10.7. *If the sum of the mean and the variance of a binomial distribution of 5 trials is 9/5, then find the binomial distribution.*

Solution. Let the binomial distribution be

$$P(x = r) = {}^nC_r p^r q^{n-r},\ 0 \le r \le n.$$

$\therefore$ Mean $= np$ and variance $= npq$.

By the given condition,

$$np + npq = \frac{9}{5} \quad \text{and} \quad n = 5.$$

$$\Rightarrow \quad 5p + 5p(1-p) = \frac{9}{5} \quad \Rightarrow \quad 5p + 5p - 5p^2 = \frac{9}{5}$$

$$\Rightarrow \quad 25p^2 - 50p + 9 = 0 \quad \therefore \quad p = \frac{1}{5}$$

$$\therefore \quad q = 1 - p = 1 - \frac{1}{5} = \frac{4}{5}.$$

$\therefore$ The binomial distribution is

$$\mathbf{P(x = r) = {}^5C_r \left(\frac{1}{5}\right)^r \left(\frac{4}{5}\right)^{5-r},\ 0 \le r \le 5.}$$

Example 10.8. *Is the following statement correct? "The mean and variance of a binomial distribution are respectively 6 and 9".*

Solution. Let the binomial distribution be

$$P(x = r) = {}^nC_r\, p^r q^{n-r}\, 0, \le r \le n.$$

NOTES

Now $\quad$ mean $= 6 \Rightarrow np = 6$

$\quad$ variance $= 9 \Rightarrow npq = 9$

$$\therefore \quad 6q = 9 \Rightarrow q = \frac{3}{2}.$$

This is impossible, because probability of an event can never be greater than 1.

$\therefore$ The given statement is not correct.

EXERCISE 10.2

1. Determine the binomial distribution whose mean is 5 and standard deviation is $\sqrt{2.5}$.
2. Determine the probability of 3 successes in a binomial distribution whose mean and variance are respectively 2 and $\frac{3}{2}$.
3. For a binomial distribution, the mean is 6 and the standard deviation is $\sqrt{2}$. Find the probability of getting 7 successes.
4. Is there any inconsistency in the statement. "The mean of a Binomial Distribution is 80 and S.D. is 8." If no inconsistency is found, what shall be the values of p, q and n?

Answers

1. $P(x = r) = {}^{10}C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{10-r}, 0 \le r \le 10$

2. 0.2076 $\quad$ 3. 0.2341 $\quad$ 4. $\frac{1}{5}, \frac{4}{5}, 400$

10.14. FITTING OF A BINOMIAL DISTRIBUTION

Let x be a binomial random variable of an experiment. Let probability of x successes be given by

$$P(x = r) = {}^nC_r\, p^r q^{n-r},\ 0 \le r \le n.$$

By fitting of a binomial distribution, we mean to find out the theoretical frequencies of the values of the binomial random variable $x = 0, 1, 2, \ldots\ldots, n$, when the experiment of n trials is repeated for, say, N times. The theoretical frequencies are given by

$$\mathbf{N.P(x = r) = N.\ {}^nC_r\ p^r q^{n-r},\ 0 \le r \le n.}$$

The recurrence formula can also be made use of, if desired.

NOTES

Example 10.9. *The following data are the number of seeds germinating out of 10 on damp filter for 80 sets of seeds. Fit a binomial distribution to this 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. **Calculation of Expected Frequencies**

x	Observed frequency (f)	fx
0	6	0
1	20	20
2	28	56
3	12	36
4	8	32
5	6	30
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
Total	80	174

Hence $n = 10$, $N = 80$

$$\text{Mean, } \bar{x} = \frac{\Sigma fx}{N} = \frac{174}{80} = 2.175$$

Let p denote the probability of *success* in a trial.

$\therefore$ Mean $= np \Rightarrow 10p = 2.175$

$\therefore$ $p = 0.2175$ and $q = 1 - p = 0.7825$

$\therefore$ The binomial distribution is

$$P(x = r) = {}^{10}C_r\,(0.2175)^r\,(0.7825)^{10-r},\ 0 \le r \le 10.$$

For $x = 0$, expected frequency $= 80 \times P(0)$

$$= 80 \times {}^{10}C_0 (0.2175)^0 (0.7825)^{10} = 6.8854 \doteq 7$$

For $x = 1$, expected frequency $= 80 \times P(1)$

$$= 80 \times {}^{10}C_1 (0.2175)^1 (0.7825)^9 = 19.1385 \doteq 19$$

For $x = 2$, expected frequency $= 80 \times P(2)$

$$= 80 \times {}^{10}C_2 (0.2175)^2 (0.7825)^8 = 23.9382 \doteq 24$$

For $x = 3$, expected frequency $= 80 \times P(3)$

$$= 80 \times {}^{10}C_3 (0.2175)^3 (0.7825)^7 = 17.7427 \doteq 18$$

For $x = 4$, expected frequency $= 80 \times P(4)$

$$= 80 \times {}^{10}C_4 (0.2175)^4 (0.7825)^6 = 8.6302 \doteq 8*$$

*We have approximated 8.6302 to 8 in order to keep the sum of all expected frequencies equal to 80.

NOTES

For $x = 5$, expected frequency $= 80 \times P(5)$

$= 80 \times {}^{10}C_5 (0.2175)^5 (0.7825)^5 = 2.8768 \doteq 3$

For $x = 6$, expected frequency $= 80 \times P(6)$

$= 80 \times {}^{10}C_6 (0.2175)^6 (0.7825)^4 = 0.6636 \doteq 1$

For $x = 7$, expected frequency $= 80 \times P(7)$

$= 80 \times {}^{10}C_7 (0.2175)^7 (0.7825)^3 = 0.1046 \doteq 0$

For $x = 8$, expected frequency $= 80 \times P(8)$

$= 80 \times {}^{10}C_8 (0.2175)^8 (0.7825)^2 = 0.0108 \doteq 0$

For $x = 9$, expected frequency $= 80 \times P(9)$

$= 80 \times {}^{10}C_9 (0.2175)^9 (0.7825)^1 = 0.0006 \doteq 0$

For $x = 10$, expected frequency $= 80 \times P(10)$

$= 80 \times {}^{10}C_{10} (0.2175)^{10} (0.7825)^0 = 0.000016 \doteq 0.$

∴ The expected frequencies are as given in the following table:

x	0	1	2	3	4	5	6	7	8	9	10
Exp. freq.	7	19	24	18	8	3	1	0	0	0	0

Example 10.10. *Five dice are thrown 96 times. The number of times 4 or 5 or 6 was actually thrown in the experiment is given in the following table:*

No. of dice (Each showing 4 or 5 or 6)	*0*	*1*	*2*	*3*	*4*	*5*
Observed frequency	*1*	*10*	*24*	*35*	*18*	*8*

Fit a binomial distribution assuming:

(i) the dice are perfect.

(ii) the nature of dice is not known.

Solution. Let x be the binomial random variable* of the experiment. The possible values of x are 0, 1, 2, 3, 4, 5. Let p be the probability of getting 4 or 5 or 6 in a single throw.

Here $n = 5, \quad N = 96$

∴ The expected frequency of getting r *successes*

$= N\,{}^nC_r p^r q^{n-r} = 96\,{}^5C_r p^r q^{5-r}, \; 0 \le r \le 5.$

(*i*) In this case, the dice are perfect.

∴ $p = \frac{3}{6} = \frac{1}{2}$

Also $q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}.$

∴ For $x = 0$, the expected frequency $= 96 \times {}^5C_0 (\frac{1}{2})^0 (\frac{1}{2})^5 = 3$

For $x = 1$, the expected frequency $= 96 \times {}^5C_1 (\frac{1}{2})^1 (\frac{1}{2})^4 = 15$

For $x = 2$, the expected frequency $= 96 \times {}^5C_2 (\frac{1}{2})^2 (\frac{1}{2})^3 = 30$

For $x = 3$, the expected frequency $= 96 \times {}^5C_3 (\frac{1}{2})^3 (\frac{1}{2})^2 = 30$

For $x = 4$, the expected frequency $= 96 \times {}^5C_4 (\frac{1}{2})^4 (\frac{1}{2})^1 = 15$

For $x = 5$, the expected frequency $= 96 \times {}^5C_5 (\frac{1}{2})^5 (\frac{1}{2})^0 = 3.$

* The experiment of throwing 5 dice can be considered as throwing of one die 5 times.

NOTES

(*ii*) In this case, the value of p will be estimated by using the observed frequencies. We have

$$\text{Mean} = \frac{\Sigma fx}{N} = \frac{(1\times 0)+(10\times 1)+(24\times 2)+(35\times 3)+(18\times 4)+(8\times 5)}{96} = \frac{275}{96}$$

Also, mean $= np$

$$\therefore \quad 5p = \frac{275}{96} \quad \text{or} \quad p = 0.5729$$

$$\therefore \quad q = 1 - p = 1 - 0.5729 = 0.4271.$$

$\therefore$ Expected frequency for $x = r$

$$= 96\ {}^5C_r\ (0.5729)^r\ (0.4271)^{5-r},\ 0 \le r \le 5.$$

For $x = 0$, the expected frequency

$$= 96 \times {}^5C_0\ (0.5729)^0\ (0.4271)^5 = 1.3643 \doteq \mathbf{1}$$

For $x = 1$, the expected frequency

$$= 96 \times {}^5C_1\ (0.5729)^1\ (0.4271)^4 = 9.1503 \doteq \mathbf{9}$$

For $x = 2$, the expected frequency

$$= 96 \times {}^5C_2\ (0.5729)^2\ (0.4271)^3 = 24.5479 \doteq \mathbf{25}$$

For $x = 3$, the expected frequency

$$= 96 \times {}^5C_3\ (0.5729)^3\ (0.4271)^2 = 32.9281 \doteq \mathbf{33}$$

For $x = 4$, the expected frequency

$$= 96 \times {}^5C_4\ (0.5729)^4\ (0.4271)^1 = 22.0844 \doteq \mathbf{22}$$

For $x = 5$, the expected frequency

$$= 96 \times {}^5C_5\ (0.5729)^5\ (0.4271)^0 = 5.9247 \doteq \mathbf{6}.$$

EXERCISE 10.3

1. 8 unbiased coins were tossed 256 times. Fit a binomial distribution.
2. 7 coins are tossed and the number of the heads noted. The experiment is repeated 128 times and the following distribution is obtained:

No. of heads	0	1	2	3	4	5	6	7
Frequency	7	16	19	35	30	13	7	1

Fit a binomial distribution assuming:

(*i*) the coins are unbiased.

(*ii*) the nature of coins is not known.

3. Four dice are thrown 162 times. The number of times 5 or 6 was actually thrown in the experiment is given in the following table:

No. of dice (Each showing 5 or 6)	0	1	2	3	4
Observed frequency	22	50	50	35	5

Fit a binomial distribution assuming:

(*i*) the dice are perfect.

(*ii*) the nature of dice is not known.

4. Ten coins were tossed 1024 times and the following frequencies are observed.

No. of heads	0	1	2	3	4	5	6	7	8	9	10
Frequency	2	10	38	106	188	257	226	128	59	7	3

Compare these frequencies with the expected frequencies.

NOTES

Answers

1. 1, 8, 28, 56, 70, 56, 28, 8, 1.
2. (*i*) 1, 7, 21, 35, 35, 21, 7, 1. (*ii*) 1, 8, 23, 36, 34, 19, 6, 1.
3. (*i*) 32, 64, 48, 16, 2 (*ii*) 18, 52, 58, 29, 5.
4. 1 (= $1024 {}^{10}C_0 (0.5135)^0 (0.4865)^{10}$), 8, 38, 107, 200, 251, 221, 133, 52, 12, 1.

III. POISSON DISTRIBUTION

10.15. INTRODUCTION

The Poisson distribution is also a discrete probability distribution. This was discovered by French mathematician **Simon Denis Poisson** (1781 – 1840) in the year 1837. This distribution deals with the evaluation of probabilities of *rare* events such as "no. of car accidents on road", "no. of earthquakes in a year", "no. of misprints in a book", etc.

10.16. CONDITIONS

The Poisson distribution is derived as a limiting case of binomial distribution. So, the conditions for the applicability of the Poisson distribution are same as those for the applicability of Binomial distribution. Here the additional requirement is that the probability of 'success' is quite near to zero.

10.17. POISSON VARIABLE

A random variable which counts the number of successes in a random experiment with trials satisfying above conditions is called a **Poisson variable.** If the probability of an article being defective is 1/500 and the event of getting a defective article is *success* and samples of 10 articles are checked for defective articles, then the possible values of the Poisson variable are 0, 1, 2, ... 10.

10.18. POISSON PROBABILITY FUNCTION

Let a random experiment satisfying the conditions of Poisson Distribution be performed. Let the number of trials in the experiment be n, which is very large. Let p denote the probability of *success* in any trial. We assume that p is very small, *i.e.*, we are dealing with a rare event. Let x denote the Poisson variable corresponding to this experiment.

$\therefore$ The possible values of x are 0, 1, 2, ..., n.

NOTES

The Poisson distribution is obtained as a limiting case of the corresponding binomial distribution of the experiment.

It can be proved mathematically that the probability of r successes is given by

$$P(x = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \ldots$$

This is called the **Poisson probability function.** The corresponding **Poisson distribution** is

x	0	1	2	3 .....
$P(x)$	$\frac{e^{-m}m^0}{0!}$	$\frac{e^{-m}m^1}{1!}$	$\frac{e^{-m}m^2}{2!}$	$\frac{e^{-m}m^3}{3!}$ .....

The constant m is the product of n and p and is called the **parameter** of the Poisson distribution.

10.19. POISSON FREQUENCY DISTRIBUTION

If a random experiment, satisfying the requirements of Poisson distribution, is repeated N times, then the expected frequency of getting r successes is given by

$$N \cdot P(x = r) = N \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \ldots$$

Remark 1. The distribution to be used in solving a problem is generally given in the problem. If it is not given, then the student should make use of Poisson distribution only when the event in the problem is of rare nature i.e., the probability of happening of event is quite near to zero.

Remark 2. The values of e^{-m} required in any particular problem is generally given with the problem itself, otherwise, the value of e^{-m} can be found out by using the table given at the end. In the examination hall, generally the table of e^{-m} is available for students. If at all the value of e^{-m} is neither given with the problem nor the table of e^{-m} is supplied in the examination hall, then the students are advised to retain their final result in terms of e^{-m}.

Example 10.11. *A company makes electric toys. The probability that an electric toy is defective is 0.01. What is the probability that a shipment of 300 toys will contain exactly 5 defective toys?*

Solution. Let x be the Poisson variable, '*no. of defective toys*'

By **Poisson distribution**, $P(x = r) = \frac{e^{-m} m^r}{r!}$, $r = 0, 1, 2, \ldots$

Here $n = 300, \quad p = 0.01$

$\therefore \quad m = np = 300 \times 0.01 = 3$

$\therefore \quad P(x = r) = \frac{e^{-3} 3^r}{r!}, \quad r = 0, 1, 2, \ldots, 300$

$P(5 \text{ defective toys}) = P(x = 5)$

$$= \frac{e^{-3}(3)^5}{5!} = \frac{0.04979 \times 243}{120} = 0.1008.$$

NOTES

Example 10.12. *10% of the tools produced in a certain factory turns out to be defective. Find the probability that in a sample of 10 tools chosen at random, (i) exactly two (ii) more than two will be defective by using Poisson approximation to binomial distribution.*

Solution. Let x be the Poisson variable, "*no. of defective tools in the sample*".

By **Poisson distribution,** $P(x = r) = \dfrac{e^{-m} m^r}{r!}$, $r = 0, 1, 2, ...$

Here $n = 10, p = \dfrac{10}{100} = \dfrac{1}{10}$ $\therefore \quad m = np = 10 \times \dfrac{1}{10} = 1$

$\therefore \quad P(x = r) = \dfrac{e^{-1}(1)^r}{r!}$, $r = 0, 1, 2, ...10.$

(*i*) P(exactly 2 defectives) = $P(x = 2)$

$$= \frac{e^{-1}}{2!} = \frac{0.36788}{2} = \mathbf{0.18394.}$$

(*ii*) P(more than 2 defectives) = $P(x > 2) = 1 - P(x \le 2)$

$$= 1 - P(x = 0 \text{ or } x = 1 \text{ or } x = 2)$$

$$= 1 - \{P(x = 0) + P(x = 1) + P(x = 2)\}$$

$$= 1 - \left\{\frac{e^{-1}}{0!} + \frac{e^{-1}}{1!} + \frac{e^{-1}}{2!}\right\} = 1 - e^{-1}\left\{1 + 1 + \frac{1}{2}\right\}$$

$$= 1 - (0.36788)(2.5) = \mathbf{0.0803.}$$

Example 10.13. *A telephone exchange receives on an average 4 calls per minute. Find the probability on the basis of Poisson distribution (m = 4), of:*

(i) 2 or less calls per minute,

(ii) upto 4 calls per minute,

(iii) more than 4 calls per minute.

Solution. Let x be the Poisson variable "*no. of calls per minute*".

By **Poisson distribution,**

$$P(x = r) = \frac{e^{-m} m^r}{r!}, \; r = 0, 1, 2, ...$$

Here m = average number of successes *i.e.*, calls per minute = 4

$$\therefore \quad P(x = r) = \frac{e^{-4}(4)^r}{r!}, \; r = 0, 1, 2, ...$$

(*i*) P(2 or less calls per minute) = $P(x \le 2)$

$$= P(x = 0 \text{ or } x = 1 \text{ or } x = 2)$$

$$= P(x = 0) + P(x = 1) + P(x = 2)$$

$$= \frac{e^{-4}(4)^0}{0!} + \frac{e^{-4}(4)^1}{1!} + \frac{e^{-4}(4)^2}{2!}$$

$$= e^{-4}\{1 + 4 + 8\} = 0.01832 \times 13 = \mathbf{0.2382}$$

NOTES

(*ii*) P(upto 4 calls per minute) = $P(x \le 4)$

$= P(x = 0 \text{ or } x = 1 \text{ or } x = 2 \text{ or } x = 3 \text{ or } x = 4)$

$= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)$

$$= \frac{e^{-4}(4)^0}{0!} + \frac{e^{-4}(4)^1}{1!} + \frac{e^{-4}(4)^2}{2!} + \frac{e^{-4}(4)^3}{3!} + \frac{e^{-4}(4)^4}{4!}$$

$$= e^{-4}\left\{1 + 4 + 8 + \frac{64}{6} + \frac{256}{24}\right\}$$

$= 0.01832 \times 34.3333 = \mathbf{0.6289}$

(*iii*) P(more than 4 calls per minute) = $P(x > 4)$

$= 1 - P(x \le 4) = 1 - 0.6289 = \mathbf{0.3711}$. [By part (*ii*)]

Example 10.14. *A manufacturer of bulbs knows that on an average 5% of his production is defective. He sells bulbs in boxes of 100 pieces and guarantees that not more than 4 bulbs will be defective in a box. What is the probability that a box will meet the guarantee ($e^{-5} = 0.0067$)?*

Solution. Let x be the Poisson variable, '*no. of defective bulbs per box*'.

By **Poisson distribution,**

$$P(x = r) = \frac{e^{-m}m^r}{r!}, \quad r = 0, 1, 2, \ldots\ldots$$

Here $n = 100, \; p = \frac{5}{100}$.

$$\therefore \quad m = np = 100 \times \frac{5}{100} = 5$$

$$\therefore \quad P(x = r) = \frac{e^{-5}5^r}{r!}, \quad r = 0, 1, 2, \ldots\ldots, 100.$$

P(box will meet the guarantee) = $P(x \le 4)$

$= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)$

$$= \frac{e^{-5}(5)^0}{0!} + \frac{e^{-5}(5)^1}{1!} + \frac{e^{-5}(5)^2}{2!} + \frac{e^{-5}(5)^3}{3!} + \frac{e^{-5}(5)^4}{4!}$$

$$= e^{-5}\left[\frac{1}{1} + \frac{5}{1} + \frac{25}{2} + \frac{125}{6} + \frac{625}{24}\right] = 0.0067 \times 65.37499 = \mathbf{0.438}.$$

Example 10.15. *A car-hire firm has two cars, which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused ($e^{-1.5} = 0.2231$).*

Solution. Let x be the Poisson variable, '*no. of demands per day*'.

$\therefore$ By **Poisson distribution,**

$$P(x = r) = \frac{e^{-m}\, m^r}{r!}, \quad r = 0, 1, 2, \ldots\ldots$$

Here m = average number of demands per day = 1.5

$$\therefore \quad P(x = r) = \frac{e^{-1.5}(1.5)^r}{r!}, \; r = 0, 1, 2, \ldots\ldots$$

NOTES

Now, proportion of days on which neither car is used

$$= P(x = 0) = \frac{e^{-1.5}(1.5)^0}{0!} = 0.2231$$

Also, proportion of days on which some demand is refused

$$= P(x > 2) = 1 - P(x \le 2) = 1 - P(x = 0 \text{ or } x = 1 \text{ or } x = 2)$$

$$= 1 - \{(P(x = 0 + P(x = 1) + P(x = 2))\}$$

$$= 1 - \left\{\frac{e^{-1.5}(1.5)^0}{0!} + \frac{e^{-1.5}(1.5)^1}{1!} + \frac{e^{-1.5}(1.5)^2}{2!}\right\}$$

$$= 1 - e^{-1.5}\left\{1 + 1.5 + \frac{2.25}{2}\right\} = 1 - (0.2231)(3.625) = \mathbf{0.1913.}$$

Example 10.16. *250 passengers have made reservations for a flight from Delhi to Mumbai. If the probability that a passenger, who has reservation, will not turn up is 0.016. Find the probability that at the most 3 passengers will not turn up.*

(given $e^{-4} = 0.0183$)

Solution. Let x be the random variable *'no. of passengers not turning up'.*

$\therefore$ By **Poisson distribution,**

$$P(x = r) = \frac{e^{-m}\, m^r}{r!}, \quad r = 0, 1, 2, ...$$

Here $n = 250, p = 0.016$

$$\therefore \quad m = np = 250 \times \frac{16}{1000} = 4$$

$$\therefore \quad P(x = r) = \frac{e^{-4}(4)^r}{r!}, r = 0, 1, 2, ..., 250$$

$\therefore$ Prob. that at most 3 passengers will not turn up

$$= P(x \le 3) = P(x = 0 \text{ or } x = 1 \text{ or } x = 2 \text{ or } x = 3)$$

$$= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$

$$= \frac{e^{-4}(4)^0}{0!} + \frac{e^{-4}(4)^1}{1!} + \frac{e^{-4}(4)^2}{2!} + \frac{e^{-4}(4)^3}{3!}$$

$$= e^{-4}\left[1 + 4 + \frac{16}{2} + \frac{64}{6}\right] = 0.0183 \times 23.67 = \mathbf{0.433.}$$

EXERCISE 10.4

1. Find the probability that at most 5 defective bolts will be found in a box of 200 bolts if it is known that 2% of such bolts are expected to be defective (you may take the distribution to be of Poisson type). ($e^{-4} = 0.0183$).
2. A telephone exchange receives on an average 3 calls per minute. Find the probability on the basis of Poisson distribution ($m = 3$), of:
   (*i*) exactly 1 call per minute (*ii*) exactly 3 calls per minute
   (*iii*) less than 3 calls per minute (*iv*) more than 1 call per minute.
3. Assuming that the probability of a total accident in a factory during a year is 1/1200, calculate the probability that in a factory employing 300 workers, there will be at least two total accidents in a year. ($e^{-0.25} = 0.7788$).

NOTES

4. The probabilities of a Poisson variate taking the values 3 and 4 are equal. Calculate the probabilities of the variate taking the values 0 and 1.
5. Find the probability that at most 5 defective articles will be found in a box of 200 articles, if experience shows that 2% of such articles are defective.
6. Assume that the probability of an individual coal miner killed in a mine accident during a year is 1/2500. Calculate the probability that in a mine employing 2000 miners, there will be (*i*) no fatal accident and (*ii*) at least one fatal accident in a year.
7. The probability that a man aged 60 years will die within a year is 0.01125. What is the probability that out of 12 such men, at least 11 will reach their 61st birth day?
8. An office switch board receives telephone calls at the rate of 3 per minute on an average. What is the probability of receiving (*i*) no call in one minute interval and (*ii*) at the most 3 calls in one minute interval?
9. 2% bulbs, manufactured by a company are defective. Find the probability that in a sample of 2000 bulbs: (*i*) less than 2 bulbs are defictive (*ii*) more than 3 bulbs are defective. (Use $e^{-4} = 0.0183$)
10. In a town 10 accidents took place in a span of 50 days. Assuming that the number of accidents per day follows the Poisson distribution, find the probability that there will be three or more accidents in a day. (use $e^{-0.2} = 0.8187$)

Answers

1. $e^{-4}\left\{\frac{4^0}{0!}+\frac{4^1}{1!}+\frac{4^2}{2!}+\frac{4^3}{3!}+\frac{4^4}{4!}+\frac{4^5}{5!}\right\} = 0.7845$ **2.** 0.1494, 0.2241, 0.4232, 0.8008
3. $1-e^{-0.25}\left(1+\frac{0.25}{1!}\right) = 0.0265$
4. $e^{-4} = 0.01832$, $4e^{-4} = 0.07328$ **5.** 0.7853
6. (*i*) 0.4493 (*ii*) 0.5507
7. P(at least 11 will survive) = P(at most one die) = $P(x \le 1) = e^{-0.135}(1 + 0.135) = 0.9916$
8. (*i*) 0.0498 (*ii*) 0.6473
9. (*i*) 0.092 (*ii*) 0.567 **10.** 0.0012.

IV. PROPERTIES OF POISSON DISTRIBUTION

10.20. THE SHAPE OF P.D.

The shape of the Poisson distribution depends upon the parameter m, the average number of successes per unit. As value of m increases, the graph of Poisson distribution would get closer to a symmetrical continuous curve.

10.21. SPECIAL USEFULNESS OF P.D.

The Poisson distribution is specially used when there are events which do not occur as outcomes of a definite number of trials in an experiment, rather occur randomly in nature. This distribution is used when the event under consideration is rare and casual. In finding probabilities by Poisson distribution, we require only the measure of average chance of occurrence (m) based on past experience or a small sample drawn for the purpose.

10.22. MEAN OF P.D.

Let x be a Poisson random variable and

$$P(x = r) = \frac{e^{-m} m^r}{r!}; \; r = 0, 1, 2, ...$$

The *mean of* x is the average numbers of successes.

$$\therefore \quad \text{Mean } (\mu) = \sum_{r=0} r.P(x = r) = \sum_{r=0} r.\frac{e^{-m} m^r}{r!}$$

$$= 0.\frac{e^{-m} m^0}{0!} + 1.\frac{e^{-m} m^1}{1!} + 2.\frac{e^{-m} m^2}{2!} + 3.\frac{e^{-m} m^3}{3!} + ...$$

$$= 0 + me^{-m}\left(\frac{1}{1!} + \frac{2m}{2!} + \frac{3m^2}{3!} + ...\right) = me^{-m}\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + ...\right)$$

$$= me^{-m} . e^m = me^0 = m . 1 = m.$$

$\therefore$ Mean of $x = \mathbf{m}$.

10.23. VARIANCE AND S.D. OF P.D.

Let x be a Poisson random variable and

$$P(x = r) = \frac{e^{-m} m^r}{r!}; \quad r = 0, 1, 2, ...$$

The variance and standard deviation of x measures the dispersion of the Poisson distribution and are given by

$$\text{variance} = \sum_{r=0} r^2.P(x = r) - \mu^2 \quad \text{and} \quad \text{S.D.} = \sqrt{\sum_{r=0} r^2.P(x = r) - \mu^2}.$$

$$\text{Now} \quad \sum_{r=0} r^2.P(x = r) = \sum_{r=0} r^2.\frac{e^{-m}m^r}{r!}$$

$$= 0^2.\frac{e^{-m}m^0}{0!} + 1^2\frac{e^{-m}m^1}{1!} + 2^2\frac{e^{-m}m^2}{2!} + 3^2\frac{e^{-m}m^3}{3!} 4^2\frac{e^{-m}m^4}{4!} + ...$$

$$= 0 + me^{-m}\left(\frac{1}{1!} + \frac{2m}{1!} + \frac{3m^2}{2!} + \frac{4m^3}{3!} .....\right)$$

$$= me^{-m}\left\{\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \frac{m^3}{3!} + ...\right) + \left(\frac{m}{1!} + \frac{2m^2}{2!} + \frac{3m^3}{3!} + ...\right)\right\}$$

$$= me^{-m}\left\{e^m + m\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + ...\right)\right\} = me^{-m}\{e^m + me^m\}$$

$$= me^{-m} e^m (1 + m) = me^0 (1 + m) = m(1 + m) = m + m^2.$$

$$\therefore \quad \text{Variance} = \sum_{r=0} r^2.P(x = r) - \mu^2 = (m + m^2) - m^2 = \mathbf{m}.$$

Also, $\quad \text{S.D.} = \sqrt{\text{variance}} = \sqrt{\mathbf{m}}.$

NOTES

10.24. γ_1 AND γ_2 OF P.D.

The values of γ_1 and γ_2 for the Poisson probability function

$$P(x = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, ...$$

are given by

$$\gamma_1 = \frac{1}{\sqrt{m}} \quad \text{and} \quad \gamma_2 = \frac{1}{m}.$$

10.25. RECURRENCE FORMULA FOR P.D.

Let x be a Poisson variable and $P(x = r) = \frac{e^{-m} m^r}{r!}$, $r = 0, 1, 2, ...$

For $k \geq 0$, $P(k) = \frac{e^{-m} m^k}{k!}$ and $P(k+1) = \frac{e^{-m} m^{k+1}}{(k+1)!}$

Dividing, we get $\frac{P(k+1)}{P(k)} = \frac{e^{-m} m^{k+1}}{(k+1)!} \cdot \frac{k!}{e^{-m} m^k} = \frac{m}{k+1}$.

$$\therefore \quad \mathbf{P(k+1) = \frac{m}{k+1} P(k), \ k = 0, 1, 2, ...}$$

This is the required **recurrence formula.**

Example 10.17. *Criticise the following statement, "The mean and standard deviation of a Poisson distribution are 5 and 2 respectively".*

Solution. Let x be a Poisson variable and

$$P(x = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, ...$$

By the given condition,

$$\text{Mean} = 5, \text{S.D.} = 2$$

$\therefore$ $\text{Variance} = (2)^2 = 4.$

Now, in Poisson distribution,

$$\text{mean} = \text{variance} = m.$$

$\therefore$ $5 = 4$. This is impossible.

$\therefore$ The given statement is incorrect.

Example 10.18. *If x is a Poisson random variable such that:*

$$P(x = 2) = 9P(x = 4) + 90P(x = 6),$$

then find mean, standard deviation and γ_1.

Solution. We have $P(x = r) = \frac{e^{-m} m^r}{r!}$, $r = 0, 1, 2, ...$

where m is the average no. of occurrence of x.

By the given condition,

$$P(x = 2) = 9\,P(x = 4) + 90\,P(x = 6).$$

$$\therefore \quad \frac{e^{-m}m^2}{2!} = 9\frac{e^{-m}m^4}{4!} + 90\frac{e^{-m}m^6}{6!}$$

or $$\frac{m^2}{2} = \frac{9m^4}{24} + \frac{90m^6}{720} \quad \text{or} \quad \frac{1}{2} = \frac{3m^2}{8} + \frac{m^4}{8} \qquad (\because \ m \neq 0)$$

or $$4 = 3m^2 + m^4 \quad \text{or} \quad (m^2 + 4)(m^2 - 1) = 0$$

$\Rightarrow$ $$m^2 = -4, 1 \quad \Rightarrow \quad m^2 = 1 \ \Rightarrow \ m = 1 \quad (\because \ m^2 > 0)$$

$\therefore$ Mean = 1, standard deviation = $\sqrt{1} = 1$ and $\gamma_1 = 1/\sqrt{1} = 1$.

NOTES

EXERCISE 10.5

1. The mean of a Poisson distribution is $\sqrt{8}$, find the value of its S.D.
2. Criticise the following statement: "The mean and variance of a Poisson distribution are 4 and 2.1 respectively."
3. Comment on the following statement: " The mean and variance of a Poisson distribution are equal only if the average occurrence of the Poisson variable is ≤ 4."
4. The standard deviation of a Poisson distribution is 3. Find the probability of getting 3 successes.

Answers

1. 1.6818 4. 0.0149.

10.26. FITTING OF A POISSON DISTRIBUTION

Let x be a Poisson random variable. Let probability of x successes be given by

$$P(x = r) = \frac{e^{-m}m^r}{r!} \qquad r = 0, 1, 2, ...$$

By fitting of a Poisson distribution, we mean to find the theoretical frequencies of the values of the Poisson random variable $x = 0, 1, 2, ...$ when the experiment is repeated N times. The theoretical frequencies are given by

$$\mathbf{N \, . \, P(x = r) = N \, . \, \frac{e^{-m}m^r}{r!} \ for \ r = 0, 1, 2, ...}$$

The use of recurrence formula would be found very useful.

Example 10.19. *The distribution of typing mistakes committed by a typist is given below:*

No. of mistakes per page	*0*	*1*	*2*	*3*	*4*	*5*
No. of pages	*142*	*156*	*69*	*27*	*5*	*1*

Assuming a Poisson model, find out the expected frequencies.

NOTES

Solution. **Calculation of Expected Frequencies**

No. of mistakes per page x	*No. of pages* f	fx
0	142	0
1	156	156
2	69	138
3	27	81
4	5	20
5	1	5
Total	N = 400	400

$\therefore$ m = average of no. of mistakes per page

$$= \frac{\Sigma fx}{N} = \frac{400}{400} = 1.$$

$\therefore$ The Poisson distribution is

$$P(x = r) = \frac{e^{-m}\, m^r}{r!} = \frac{e^{-1}(1)^r}{r!} = \frac{e^{-1}}{r!}, \qquad r = 0, 1, 2, \ldots\ldots$$

For $x = 0$, expected frequency = 400 . P(0)

$$= 400 \cdot \frac{e^{-1}}{0!} = 400(0.36788) = 147.152 \doteq \mathbf{147}$$

For $x = 1$, expected frequency = 400 . P(1)

$$= 400 \cdot \frac{e^{-1}}{1!} = 400(0.36788) = 147.152 \doteq \mathbf{147}$$

For $x = 2$, expected frequency = 400 . P(2)

$$= 400 \cdot \frac{e^{-1}}{2!} = 400(0.36788)/2 = 73.576 \doteq \mathbf{74}$$

For $x = 3$, expected frequency = 400 . P(3)

$$= 400 \cdot \frac{e^{-1}}{3!} = 400(0.36788)/6 = 24.525 \doteq \mathbf{25}$$

For $x = 4$, expected frequency = 400 . P(4)

$$= 400 \cdot \frac{e^{-1}}{4!} = 400(0.36788)/24 = 6.131 \doteq \mathbf{6}$$

For $x = 5$, expected frequency = 400 . P(5)

$$= 400 \cdot \frac{e^{-1}}{5!} = 400(0.36788)/120 = 1.2263 \doteq \mathbf{1.}$$

Example 10.20. *Letters were received in an office on each of 100 days. Assuming the following data to form a random sample from a Poisson distribution, find the expected frequencies, correct to the nearest unit, taking $e^{-4} = 0.0183$.*

No. of letters	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Frequency	*1*	*4*	*15*	*22*	*21*	*20*	*8*	*6*	*2*	*0*	*1*

NOTES

Solution. **Calculation of Expected Frequencies**

No. of letters x	*Frequency* f	fx
0	1	0
1	4	4
2	15	30
3	22	66
4	21	84
5	20	100
6	8	48
7	6	42
8	2	16
9	0	0
10	1	10
Total	N = 100	400

$\therefore$ m = average no. of letters per day

$$= \frac{\Sigma fx}{N} = \frac{400}{100} = 4.$$

$\therefore$ The Poisson distribution is

$$P(x = r) = \frac{e^{-m} m^r}{r!} = \frac{e^{-4} 4^r}{r!}, \quad r = 0, 1, 2, ...$$

By recurrence formula,

$$P(k + 1) = \frac{4}{k+1} P(k), \quad k = 0, 1, 2, ...$$

Let $f(x)$ denote the expected frequency of x.

$$\therefore \qquad f(k + 1) = N \,.\, P(k + 1) = 100 \,.\, \frac{4}{k+1} P(k) = \frac{4}{k+1} 100 \,.\, P(k) = \frac{4}{k+1} f(k)$$

$$\therefore \qquad f(k + 1) = \frac{4}{k+1} f(x), \quad k = 0, 1, 2, ...$$

Now $$f(0) = 100\, P(0) = 100 \,.\, \frac{e^{-4}(4)^0}{0!} = 100(0.0183) = 1.83 \doteq 2$$

$$f(1) = \frac{4}{1} f(0) = 4(1.83) = 7.32 \doteq 7$$

$$f(2) = \frac{4}{2} f(1) = 2(7.32) = 14.64 \doteq 15$$

$$f(3) = \frac{4}{3} f(2) = \frac{4(14.64)}{3} = 19.52 \doteq 20$$

$$f(4) = \frac{4}{4} f(3) = \frac{4(19.52)}{4} = 19.52 \doteq 20$$

$$f(5) = \frac{4}{5} f(4) = \frac{4(19.52)}{5} = 15.62 \doteq 16$$

NOTES

$$f(6) = \frac{4}{6} f(5) = \frac{4(15.62)}{6} = 10.41 \doteq 10$$

$$f(7) = \frac{4}{7} f(6) = \frac{4(10.41)}{7} = 5.95 \doteq 6$$

$$f(8) = \frac{4}{8} f(7) = \frac{4(5.95)}{8} = 2.97 \doteq 3$$

$$f(9) = \frac{4}{9} f(8) = \frac{4(2.97)}{9} = 1.32 \doteq 1$$

$$f(10) = \frac{4}{10} f(9) = \frac{4(1.32)}{10} = 0.53 \doteq 1.$$

EXERCISE 10.6

1. A typist commits the following number of mistakes per page in typing 100 pages. Fit a Poisson distribution and calculate theoretical frequencies:

Mistakes per page	0	1	2	3	4	5
Frequency	42	33	14	6	4	1

You are given, $e^{-1} = 0.3679$.

2. Below are given the number of vacancies of judges occurring in a High Court over a period of 96 years:

Fit a Poisson distribution to represent the frequencies of vacancies per year and find the expected frequencies:

No. of vacancies per year	0	1	2	3
No. of years	59	27	9	1

3. In 1,000 sets of trials for an event of small frequencies f_i, of the number of x_i successes are:

x	0	1	2	3	4	5	6	7
f	305	365	210	80	28	9	2	1

Fit a Poisson distribution to the above data and calculate the theoretical frequencies.

4. 5,000 television sets are inspected as they come off the production line and the number of defects per set is recorded below.

No. of defects	0	1	2	3	4
No. of sets	3680	720	520	70	10

Estimate the average number of defects per set and the expected frequencies of 0, 1, 2, 3 and 4 defects, assuming Poisson distribution.

Answers

1. 37, 37, 18, 6, 2, 0
2. 58, 29, 7, 1
3. 301, 361, 217, 87, 26, 6, 1, 0.
4. Average no. of defect per set = 0.402 ; 3351, 1341, 268, 36, 4.

V. NORMAL DISTRIBUTION

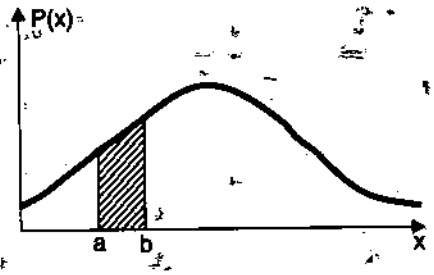
NOTES

10.27. INTRODUCTION

In Binomial and Poisson distributions, we considered the probabilities of discrete random variables. Now we shall consider random variables which may take non-countably infinitely many possible values. Such a random variable is called a **continuous random variable**. The random variables corresponding to the measurement of height, weight, etc. are continuous. We have already discussed probability distributions of discrete random variables. We shall also be considering the probability function of a very important continuous random variable, namely *normal variable*.

10.28. PROBABILITY FUNCTION OF CONTINUOUS RANDOM VARIABLE

In discrete probability distributions, the probability is defined for each and every value of the variable and the sum of all these probabilities is one. On the other hand, continuous random variables are defined over intervals of real numbers which contains non-countably infinitely many numbers. Let x be a continuous random variable. The probability of x to take any particular value is generally zero. For example, if an individual is selected at random from a large group of males, then the probability that his weight (x) is exactly 56 kg (*i.e.*, 56.000 ..... kg) would be zero. On the other hand, the probability that weight (x) lying between 55.600 ..... kg and 56.200 ..... kg need not be zero. Thus, we cannot define a probability function for a continuous random variable as we did in the case of a discrete random variable. In case of a continuous random variable (x), the probability of x taking any particular value is generally zero and practically does not make any sense whereas the probability of x taking values between any two different values is meaningful.



For a continuous random variable, x, a function $P(x)$ is called a **probability function** if:

(*i*) $P(x) \geq 0$ and (*ii*) $\int_{-\infty}^{\infty} P(x)\,dx = 1$

If $P(x)$ is a *probability function* of x, then we define:

$$P(a < x < b) = \int_a^b P(x)\,dx .$$

Thus, if $P(x)$ is a *probability function* of x, then:

(*i*) $P(x)$ is non-negative

(*ii*) area bounded by the curve and x-axis is equal to one

NOTES

(*iii*) area bounded by the curve, x-axis and ordinates $x = a, x = b$ gives the measure of the probability that x lies between a and b.

Remark. Since the probability of x taking any particular value is generally zero, we have

$$P(a < x < b) = P(a \le x < b) = P(a < x \le b) = P(a \le x \le b)$$

10.29. NORMAL DISTRIBUTION

The normal distribution is a particular type of continuous probability distribution. This was discovered by **De Moivre (1667—1754)** in the year 1733. The normal distribution is obtained as a limiting case of a binomial distribution when n, the number of trials is indefinitely large and neither p nor q is very small.

10.30. DEFINITION

A continuous random variable x is said to have a **normal distribution** (N.D.) if its probability function is given by

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty, \qquad ...(1)$$

where μ and σ are the mean and standard deviation of the distribution respectively.

Remark. If x is a normal variable with mean μ and variance σ^2, then we write symbolically as $x \sim N(\mu, \sigma^2)$.

10.31. STANDARD NORMAL DISTRIBUTION

Let x be a normal variable with probability function:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

where μ and σ are the mean and standard deviation of the distribution respectively.

We define $z = \dfrac{x-\mu}{\sigma}$.

It can be proved mathematically that z is also a normal variable with mean zero and variance one. A normal variable with mean zero and variance one is called a **standard normal variable (S.N.V.).**

In terms of z, the probability function of x reduces to

$$P(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty$$

Remark. Thus, if $x \sim N(\mu, \sigma^2)$ and z = S.N.V. of $x = \dfrac{x-\mu}{\sigma}$, then $z \sim N(0, 1)$.

10.32. AREA UNDER NORMAL CURVE

NOTES

Let x be a *normal variable* with mean μ and standard deviation σ. Let $z = \frac{x-\mu}{\sigma}$ be the corresponding S.N.V. We know that mean and standard deviation of the variable z are 0 and 1 respectively. Therefore, the curves of standard normal variables corresponding of normal variables are identical.

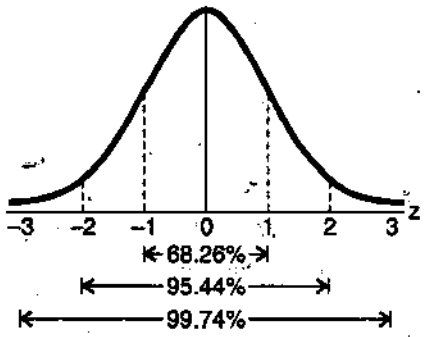
This is why the curve of a standard normal variable is called the **standard normal curve**. For the standard normal curve, we have:

(*i*) area between $z = -1$ and $z = 1$ is 68.26% of total area, which is one.

$\therefore$ $P(-1 \le z \le 1) = 0.6826$.

(*ii*) area between $z = -2$ and $z = 2$ is 95.44% of total area.

$\therefore$ $P(-2 \le z \le 2) = 0.9544$.



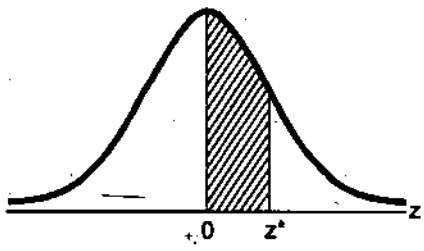
(*iii*) area between $z = -3$ and $z = 3$ is 99.74% of total area.

$\therefore$ $P(-3 \le z \le 3) = 0.9974$.

The area bounded by the curve of a S.N.V. z, z-axis and the ordinates at $z = 0$ and any positive value of z is provided by a standard table. This knowledge of measure of area in case of S.N.V. is used to find the area bounded by the corresponding normal variable x, x-axis and any two ordinates. This in turn would help us to find the probability of normal variable x lying between any two real numbers.

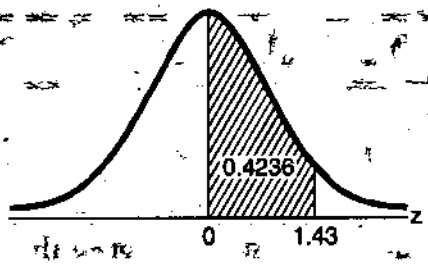
10.33. TABLE OF AREA UNDER STANDARD NORMAL CURVE

The table titled *area under standard normal curve* is given at the end. Let z^* be any arbitrary but fixed value of the variable z. The first column of the table provides for z values with one decimal digit and the second column gives areas bounded between z-curve and ordinates $z = 0$ and $z = z^*$, which is equal to $P(0 \le z \le z^*)$.



NOTES

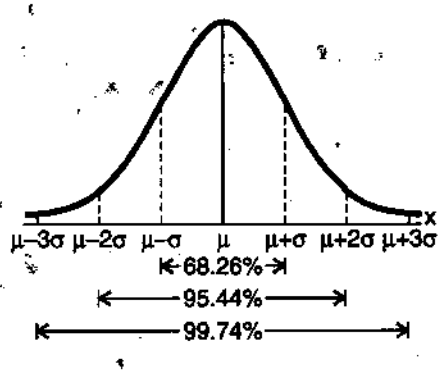
For example, from the table $P(0 \le z \le 1.4) = 0.4192$. The first row of the table provides for the second decimal digit of z^*. For example, $P(0 \le z \le 1.43) = 0.4236$.



Remark. Sometimes, the table for the probabilities $P(-\infty < z \le z^*)$ is given in the examination hall. In such a case, the students should find the value of $P(0 \le z \le z^*)$ by using the following formula:

$$P(0 \le z \le z^*) = P(-\infty < z \le z^*) - 0.5.$$

10.34. PROPERTIES OF NORMAL DISTRIBUTION



(*i*) The area bounded by the curve and the x-axis is equal to one.

(*ii*) $P(a < x < b)$, *i.e.* the probability that x lies between a and b is equal to the area bounded by the curve, x-axis and ordinates $x = a$ and $x = b$.

(*iii*) The curve is bell-shaped and symmetrical about the line $x = \mu$.

(*iv*) The mean (μ) and variance (σ^2) of a normal distribution are called its *parameters*.

(*v*) The location and shape of a normal distribution depends upon the values of its parameters.

(*vi*) If the mean and S.D. of a normal distribution are μ and σ respectively, then:

(*a*) $P(\mu - \sigma \le x \le \mu + \sigma) = P\left(\frac{\mu - \sigma - \mu}{\sigma} \le \frac{x - \mu}{\sigma} \le \frac{\mu + \sigma - \mu}{\sigma}\right)$

$= P(-1 \le z \le 1) = 0.6826$

(*b*) $P(\mu - 2\sigma \le x \le \mu + 2\sigma) = P\left(\frac{\mu - 2\sigma - \mu}{\sigma} \le \frac{x - \mu}{\sigma} \le \frac{\mu + 2\sigma - \mu}{\sigma}\right)$

$= P(-2 \le z \le 2) = 0.9544$

(*c*) $P(\mu - 3\sigma \le x \le \mu + 3\sigma) = P\left(\frac{\mu - 3\sigma - \mu}{\sigma} \le \frac{x - \mu}{\sigma} \le \frac{\mu + 3\sigma - \mu}{\sigma}\right)$

$= P(-3 \le z \le 3) = 0.9974.$

(*vii*) The mean, median and mode of a normal distribution coincide with each other.

(*viii*) For a normal distribution, $\gamma_1 = 0$ and $\gamma_2 = 0$. Therefore, normal distribution is *mesokurtic*.

(*ix*) In a normal distribution, mean deviation about mean is approximately equal to $\frac{4}{5}$ time its standard deviation.

(*x*) In a normal distribution, quartile deviation is approximately equal to $\frac{2}{3}$ times its standard deviation.

(*xi*) The tails of the curve of a normal distribution extend indefinitely on both sides of $x = \mu$ and never touches the x-axis.

(*xii*) In a normal distribution, Q_1 and Q_3 are equidistant from the median.

NOTES

Example 10.21. *The mean and standard deviation of a normal variable x are 50 and 4 respectively. Find the values of the corresponding standard normal variable, when x is equal to 42, 84, 85, 32 and 40.*

Solution. We have $\mu = 50, \sigma = 4$.

Let z be the standard normal variable corresponding to x.

$$\therefore \quad z = \frac{x - \mu}{\sigma} = \frac{x - 50}{4}$$

$\therefore$ When $x = 42$, $z = \frac{42 - 50}{4} = -2$

When $x = 84$, $z = \frac{84 - 50}{4} = 8.5$

When $x = 85$, $z = \frac{85 - 50}{4} = 8.75$

When $x = 32$, $z = \frac{32 - 50}{4} = -4.5$

When $x = 40$, $z = \frac{40 - 50}{4} = -2.5$.

Example 10.22. *Find the area under the standard normal curve which lies:*

(i) to the right of z = 2.70 *(ii) to the left of z = 1.73*

(iii) to the right of z = – 0.66 *(iv) to the left of z = – 1.88*

(v) between z = 1.25 and z = 1.67

(vi) between z = – 1.85 and z = – 0.90

(vii) between z = – 1.45 and z = 1.45

(viii) between z = – 0.9 and z = 1.58.

Solution. The variable z is a standard normal variable (S.N.V.).

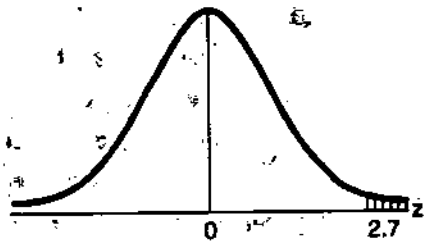
$\therefore$ Total area under the curve of z and z-axis is one.

(*i*) Area to the right of $z = 2.7$

$= 0.5 -$ area between $z = 0$ and $z = 2.7$

$= 0.5 - 0.4965$ (Using area Table)

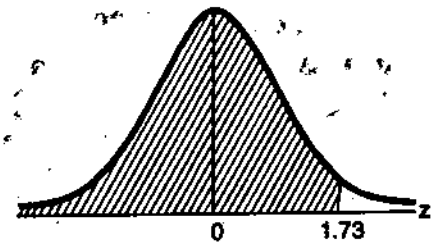
$= \mathbf{0.0035}$.



(*ii*) Area to the left of $z = 1.73$

$= 0.5 +$ area between $z = 0$ and $z = 1.73$

$= 0.5 + 0.4582 = \mathbf{0.9582}$.

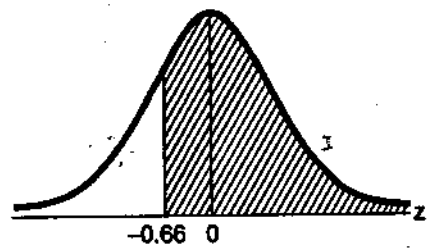


(*iii*) Area to the right of $z = -0.66$

$=$ (area from $z = -0.66$ to $z = 0$) $+ 0.5$

$=$ (area from $z = 0$ to $z = 0.66$) $+ 0.5$ (By symmetry of curve)

$= 0.2454 + 0.5 = \mathbf{0.7454}$. (Using Table)



NOTES

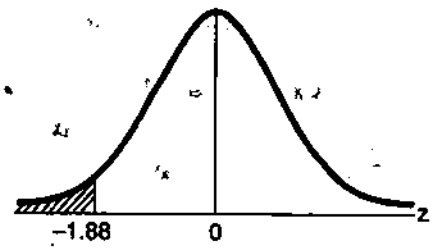
(*iv*) Area to the left of $z = -1.88$

$= 0.5 -$ area between $z = -1.88$ and $z = 0$

$= 0.5 -$ area between $z = 0$ and $z = 1.88$ (By symmetry of curve)

$= 0.5 - 0.4699 =$ **0.0301.**

(Using Table)

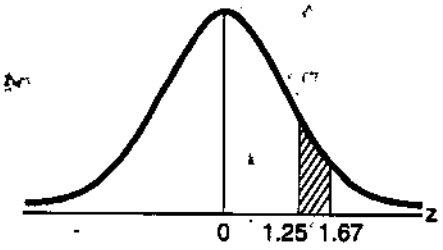


(*v*) Area between $z = 1.25$ and $z = 1.67$

= (area between $z = 0$ and $z = 1.67$) – (area between $z = 0$ and $z = 1.25$)

$= 0.4525 - 0.3944 =$ **0.0581.**

(Using Table)



(*vi*) Area between $z = -1.85$ and $z = -0.90$

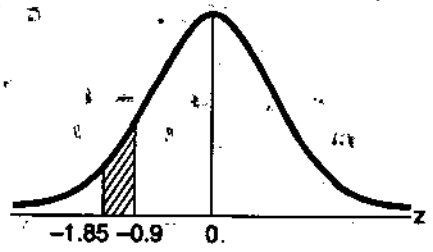
= (area between $z = -1.85$ and $z = 0$) – (area between $z = -0.90$ and $z = 0$)

= (area between $z = 0$ and $z = 1.85$) – (area between $z = 0$ and $z = 0.90$)

(By symmetry of curve)

$= 0.4678 - 0.3159 =$ **0.1519.**

(Using Table)



(*vii*) Area between $z = -1.45$ and $z = 1.45$

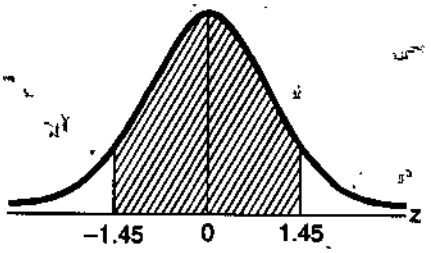
= (area between $z = -1.45$ and $z = 0$) + (area between $z = 0$ and $z = 1.45$)

= (area between $z = 0$ and $z = 1.45$) + (area between $z = 0$ and $z = 1.45$)

(By symmetry of curve)

$= 2$(area between $z = 0$ and $z = 1.45$)

$= 2(0.4265) =$ **0.8530.** (Using Table)



(*viii*) Area between $z = -0.9$ and $z = 1.58$

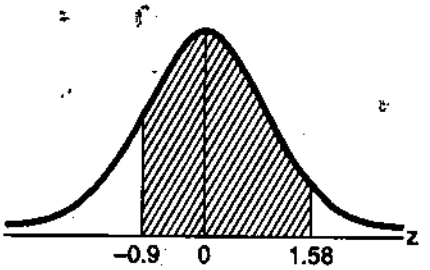
= (area between $z = -0.9$ and $z = 0$) + (area between $z = 0$ and $z = 1.58$)

= (area between $z = 0$ and $z = 0.9$) + (area between $z = 0$ and $z = 1.58$)

(By symmetry of curve)

$= 0.3159 + 0.4429 =$ **0.7588.**

(Using Table)



Example 10.23. *x is a normal variable with mean 25 and standard deviation 5. Find the probability that :*

(i) $x \leq 10$ *(ii)* $15 \leq x \leq 30$

(iii) $| x - 30 | \geq 10$.

NOTES

Solution. We have mean = 25, S.D. = 5.

Let z be the S.N.V. corresponding to x.

$$\therefore \quad z = \frac{x-\mu}{\sigma} = \frac{x-25}{5}.$$

(*i*) When $x = 10$, $z = \dfrac{10-25}{5} = -3$

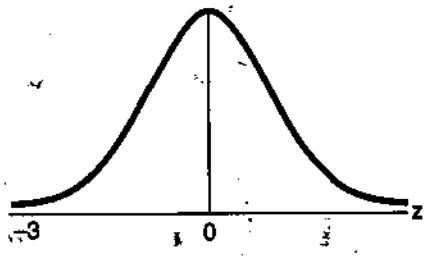
$\therefore$ Required probability $= P(x \le 10)$

$= P(z \le -3) = P(z \ge 3)$

$= 0.5 - P(0 \le z \le 3)$

$= 0.5 - 0.4987$

$= \mathbf{0.0013.}$



(*ii*) When $x = 15$, $z = \dfrac{15-25}{5} = -2$

When $x = 30$, $z = \dfrac{30-25}{5} = 1$

$\therefore$ Required probability

$= P(15 \le x \le 30)$

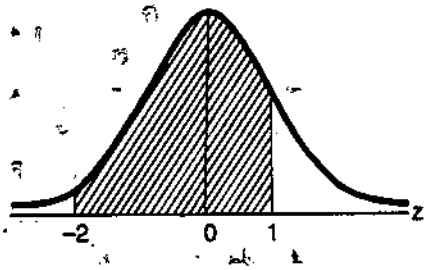
$= P(-2 \le z \le 1)$

$= P(-2 \le z \le 0) + P(0 \le z \le 1)$

$= P(0 \le z \le 2) + P(0 \le z \le 1)$

$= 0.4772 + 0.3413$

$= \mathbf{0.8185.}$



(*iii*) Required probability $= P(|x - 30| \ge 10)$

$= 1 - P(|x - 30| < 10)$

$= 1 - P(30 - 10 < x < 30 + 10)$

$= 1 - P(20 < x < 40)$

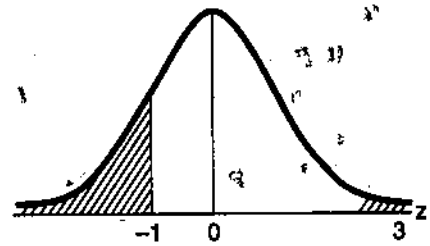
$$= 1 - P\left(\frac{20-25}{5} < \frac{x-25}{5} < \frac{40-25}{5}\right)$$

$= 1 - P(-1 < z < 3)$

$= 1 - \{P(-1 \le z \le 0) + P(0 \le z \le 3)\}$ $\quad [\because \ P(z = -1) = P(z = 3) = 0]$

$= 1 - \{P(0 \le z \le 1) + P(0 \le z \le 3)\}$

$= 1 - \{0.3413 + 0.4987\} = \mathbf{0.16.}$



Example 10.24. *In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.*

Solution. Let x be the normal variable and z be its S.N.V. Let μ and σ be the mean and standard deviation of x.

By the given conditions,

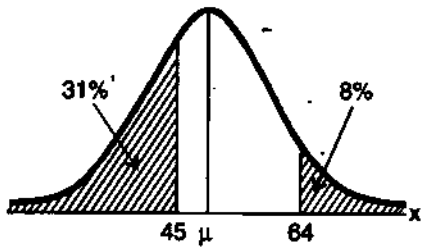
$$P(x < 45) = \frac{31}{100} = 0.31 \quad \text{and} \quad P(x > 64) = \frac{8}{100} = 0.08.$$

Since $P(x < 45) < 0.5$. $\therefore$ $x = 45$ lies on the left of $x = \mu$ and so the value of the corresponding z-variable is –ve.

NOTES

When $x = 45$, let $z = \dfrac{45-\mu}{\sigma} = -z_1\ (z_1 > 0)$

When $x = 64$, let $z = \dfrac{64-\mu}{\sigma} = z_2\ (> 0)$



$\therefore \quad P(x < 45) = 0.31$

$\Rightarrow \quad P(z < -z_1) = 0.31$

$\Rightarrow \quad P(z > z_1) = 0.31$

$\Rightarrow \quad 0.5 - P(0 \le z \le z_1) = 0.31$

$\Rightarrow \quad P(0 \le z \le z_1) = 0.19$

$\Rightarrow \quad z_1 = 0.5$ (From area table)

$\therefore \quad \dfrac{45-\mu}{\sigma} = -0.5 \quad$ or $\quad 45 - \mu = -0.5\,\sigma \quad ...(1)$

Also $\quad P(x > 64) = 0.08$

$\Rightarrow \quad P(z > z_2) = 0.08 \quad \Rightarrow \quad 0.5 - P(0 \le z \le z_2) = 0.08$

$\Rightarrow \quad P(0 \le z \le z_2) = 0.42 \quad \Rightarrow \quad z_2 = 1.4$

(From area table)

$\therefore \quad \dfrac{64-\mu}{\sigma} = 1.4 \quad$ or $\quad 64 - \mu = 1.4\,\sigma \quad ...(2)$

(1) – (2) $\Rightarrow \quad -19 = -1.9\sigma \quad \Rightarrow \quad \sigma = 10$

$\therefore$ (1) $\Rightarrow \quad 45 - \mu = -0.5\,(10) = -5$

$\Rightarrow \quad \mu = 45 + 5 = 50$

$\therefore \quad \mu = 50$ and $\sigma = 10$.

Example 10.25. *The profits of 400 companies are normally distributed with mean ₹ 150 lakhs and standard deviation ₹ 20 lakhs. Estimate the number of companies with:*

(i) profits less than ₹ 128 lakhs

(ii) profits more then ₹ 175 lakhs

(iii) profits between ₹ 100 lakhs and ₹ 138 lakhs.

Solution. Let x be the normal variable *'profit'*. Let z be the corresponding S.N.V.

$\therefore \quad z = \dfrac{x-\mu}{\sigma} = \dfrac{x-150}{20}$

(i) P(profit less than ₹ 128 lakhs) = $P(x < 128)$

$= P(x - 150 < 128 - 150)$

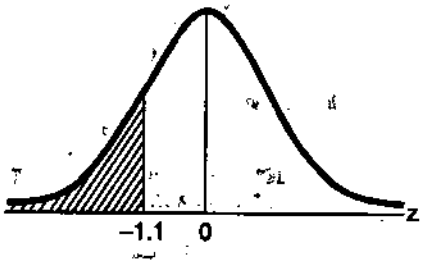
$= P\left(\dfrac{x-150}{20} < \dfrac{-22}{20}\right)$

$= P(z < -1.1) = P(z > 1.1)$

(By symmetry)

$= 0.5 - P(0 < z \le 1.1)$

$= 0.5 - 0.3643 = 0.1357.$



No. of companies with profit less than ₹ 128 lakhs

$= 400\ P(x < 128) = 400 \times 0.1357 \doteq 54.$

NOTES

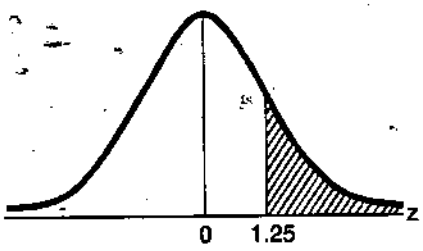
(*ii*) P(profit more than ₹ 175 lakhs)

$= P(x > 175) = P(x - 150 > 175 - 150)$

$= P\left(\frac{x-150}{20} > \frac{25}{20}\right) = P(z \geq 1.25)$

$= 0.5 - P(0 < z \leq 1.25)$

$= 0.5 - 0.3944 = 0.1056.$



∴ No. of companies with profit more than ₹ 175 lakhs

$= 400\ P(x > 175)$

$= 400 \times 0.1056 = 42.$

(*iii*) P(profit between ₹ 100 lakhs and ₹ 138 lakhs)

$= P(100 < x < 138)$

$= P(100 - 150 < x - 150 < 138 - 150)$

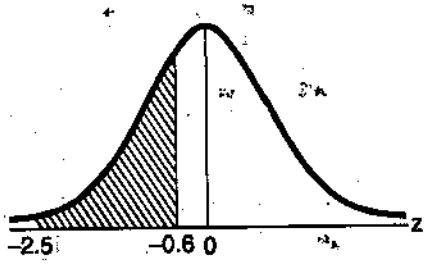
$= P\left(\frac{-50}{20} < \frac{x-150}{20} < \frac{-12}{20}\right)$

$= P(-2.5 < z < -0.6)$

$= P(0.6 < z < 2.5)$ (By symmetry)

$= P(0 < z < 2.5) - P(0 < z < 0.6)$

$= 0.4938 - 0.2257 = 0.2681.$



∴ No. of companies with profits between ₹ 100 lakhs and ₹ 138 lakhs

$= 400\ P(100 < x < 138)$

$= 400 \times 0.2681 = 107.$

Example 10.26. *The marks obtained by students in a degree examination are normally distributed. The mean marks and S.D. of the distribution are 500 and 100 respectively. If 674 appeared in the examination and out of these, 550 are to be declared passed, what should be the minimum pass marks?*

Solution. Let x denote the normal variable *marks*. Let z be the S.N.V. of x.

$$z = \frac{x - 500}{100}$$

Let minimum pass marks be k.

$\therefore \quad 674\ P(x \geq k) = 550$

$\Rightarrow \quad P(x \geq k) = \frac{550}{674} = 0.816$

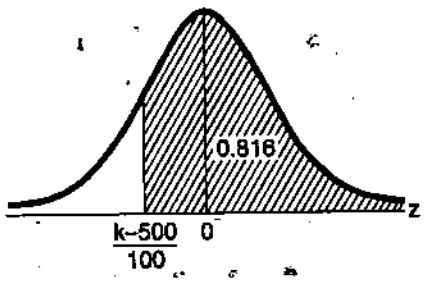
or $\quad P\left(\frac{x-500}{100} \geq \frac{k-500}{100}\right) = 0.816$

or $\quad P\left(z \geq \frac{k-500}{100}\right) = 0.816$

or $\quad P\left(\frac{k-500}{100} \leq z\right) = 0.816$

or $\quad P\left(\frac{k-500}{100} \leq z \leq 0\right) + 0.5 = 0.816$

or $\quad P\left(\frac{k-500}{100} \leq z \leq 0\right) = 0.316$



NOTES

or $P\left(0 \le z \le -\frac{k-500}{100}\right) = 0.316$ (By symmetry of normal curve)

By area table,

$P(0 \le z \le 0.9) = 0.316$ (Approx.)

$\therefore \quad -\frac{k-500}{100} = 0.9$

i.e., $500 - k = 90$ or $k = 500 - 90 = 410$.

$\therefore$ Minimum pass marks = **410.**

Example 10.27. *Assuming the mean height of soldiers to be 68.22 inches with a variance of 10.8 (inches)^2, find how many soldiers in a regiment of 10,000 would you expect to be over 6 feet tall?*

Solution. Let x denote the variable *height*. We assume that x is normally distributed. The mean and S.D. of x are 68.22 and $\sqrt{10.8} = 3.2863$.

Let z be the corresponding S.N.V.

$\therefore \quad z = \frac{x - 68.22}{3.2863}$

Now, the expected number of soldiers, who are at least 6 feet (72 inches) tall

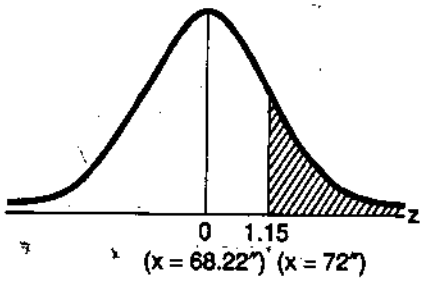
$= 10{,}000\ P(x > 72)$

$= 10{,}000\ P\left(\frac{x - 68.22}{3.2863} > \frac{72 - 68.22}{3.2863}\right)$

$= 10{,}000\ P(z > 1.15)$

$= 10{,}000\ (0.5 - P(0 \le z \le 1.15))$

$= 10{,}000\ (0.5 - 0.3749) = \mathbf{1251}.$



Example 10.28. *In a certain examination, the percentages of passes and distinctions were 45 and 9 respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75. (Assume the distribution of marks to be normal).*

Solution. Let x denote the normal variable *marks*. Let μ and σ be the mean and standard deviation of x respectively. Let z be the S.N.V. of x.

$\therefore \quad z = \frac{x - \mu}{\sigma}$

Now, percentage of passes = 45%

$\therefore \quad 100\ P(x \ge 40) = 45$

$\Rightarrow \quad P(x \ge 40) = 0.45 \quad \Rightarrow \quad P\left(\frac{x-\mu}{\sigma} \ge \frac{40-\mu}{\sigma}\right) = 0.45$

$\Rightarrow \quad P\left(z \ge \frac{40-\mu}{\sigma}\right) = 0.45 \quad \Rightarrow \quad 0.5 - P\left(0 \le z \le \frac{40-\mu}{\sigma}\right) = 0.45$

$\Rightarrow \quad P\left(0 \le z \le \frac{40-\mu}{\sigma}\right) = 0.05.$

Also from the table,

$P(0 \le z \le 0.12) = 0.05$

NOTES

$$\therefore \quad \frac{40-\mu}{\sigma} = 0.12 \qquad \ldots(1)$$

Also, percentage of distinctions = 9%

$$\therefore \quad 100\,P(x \geq 75) = 9 \quad \Rightarrow \quad P(x \geq 75) = 0.09$$

$$\Rightarrow \quad P\left(\frac{x-\mu}{\sigma} \geq \frac{75-\mu}{\sigma}\right) = 0.09 \quad \Rightarrow \quad P\left(z \geq \frac{75-\mu}{\sigma}\right) = 0.09$$

$$\Rightarrow \quad 0.5 - P\left(0 \geq z \geq \frac{75-\mu}{\sigma}\right) = 0.09 \Rightarrow P\left(0 \leq z \leq \frac{75-\mu}{\sigma}\right) = 0.41$$

Also, from the table,

$$P(0 \leq z \leq 1.34) = 0.41$$

$$\therefore \quad \frac{75-\mu}{\sigma} = 1.34. \qquad \ldots(2)$$

Dividing (1) by (2), we get

$$\frac{40-\mu}{75-\mu} = \frac{0.12}{1.34} \Rightarrow \mu = 36.5576 \doteq 37$$

$\therefore$ Average marks = 37.

EXERCISE 10.7

1. The mean and standard deviation of a normal variable are 35 and 5 respectively. Find the values of the corresponding S.N.V., when $x = 10, 15, 22, 34, 35, 55$!
2. If z is a standard normal variable, then find the following probabilities:
   (*i*) $P(1 \leq z \leq 2)$ (*ii*) $P(z \geq 3)$.
3. x is a normal variable with mean 50 and standard deviation 8. Find the probabilities:
   (*i*) $x \leq 60$ (*ii*) $10 \leq x \leq 40$;
   (*iii*) $x > 60$.
4. A normal curve has $\bar{x} = 40$ and $\sigma = 15$. Find the area between $x_1 = 25$ and $x_2 = 60$.
5. The income of a group of 5000 persons was found to be normally distributed with mean ₹ 700 and standard deviation ₹ 50. Find the expected number of persons getting (*i*) less than Rs. 680 (*ii*) more than ₹ 750 and (*iii*) between ₹ 680 and ₹ 750.
6. The marks obtained by a large group of students in a final examination in statistics have mean 68 and standard deviation 9. If these marks are normally distributed, what percentage of students can you expect to have secured marks between 60 and 65, both inclusive?
7. In a sample of 120 workers in a factory, the mean and standard deviation of wages were ₹ 11.35 and ₹ 3.03 respectively. Find the percentage of workers getting wages between ₹ 9 and ₹ 17 in the whole factory, assuming that the wages to be normally distributed.
8. 5000 candidates appeared in a certain examination paper carrying a maximum of 100 marks. It was found that the marks were normally distributed with mean 39.5 and standard deviation 12.5. Determine the approximate number of candidates who secured a first class for which a minimum of 60 is necessary.
9. In a large group of persons, it is found that 5% are under 60 inches and 40% are between 60 and 65 inches in height. Assuming the distribution to be normal, find the mean and standard deviation of the height.

NOTES

Answers

1. $-5, -4, -2.6, -0.2, 0, 4$ 2. (*i*) 0.1359 (*ii*) 0.0013
3. (*i*) 0.8944 (*ii*) 0.1056 (*iii*) 0.1056
4. 0.7495 5. (*i*) 1723 (*ii*) 793 (*iii*) 2484
6. 18.4% 7. 100 $P(9 < x < 17) = 75.09\%$ 8. 252
9. 65.41, 3.28.

10.35. FITTING OF A NORMAL DISTRIBUTION

Let x be a normal variable. Let the probability density function $P(x)$ of x be given by

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty,$$

where μ and σ are the mean and standard deviation of the distribution respectively. For calculating the expected normal frequencies, corresponding to an observed frequency distribution, the following steps are taken:

(*i*) The values of mean (μ) and S.D. (σ) are calculated by usual methods.

(*ii*) The values of the standard normal variable $z = \frac{x-\mu}{\sigma}$ are calculated corresponding to each lower limit of classes in the given distribution.

(*iii*) Corresponding to these values of z, the areas under the normal curve to the left of these ordinates are calculated. This is done by using the table given at the end.

(*iv*) The areas for the successive classes are obtained by subtracting the corresponding areas calculated in step (*iii*).

(*v*) The areas for the successive classes are multiplied by N to get the required expected frequencies.

Example 10.29. *Fit a normal curve to the following data:*

Class	*0—5*	*5—10*	*10—15*	*15—20*	*20—25*	*25—30*	*30—35*	*35—40*	*40—45*
No. of items	*20*	*24*	*32*	*28*	*20*	*16*	*34*	*10*	*16*

Solution. **Calculation of Mean and S.D.**

Class	*f*	*x*	$d = x - A$, $A = 22.5$	$u = d/h$, $h = 5$	*fu*	fu^2
0—5	20	2.5	−20	−4	−80	320
5—10	24	7.5	−15	−3	−72	216
10—15	32	12.5	−10	−2	−64	128
15—20	28	17.5	−5	−1	−28	28
20—25	20	22.5	0	0	0	0
25—30	16	27.5	5	1	16	16
30—35	34	32.5	10	2	68	136
35—40	10	37.5	15	3	30	90
40—45	16	42.5	20	4	64	250
Total	200				−66	1190

$$\text{Mean } (\mu) = A + \left(\frac{\Sigma fu}{N}\right) h = 22.5 + \left(\frac{-66}{200}\right) 5 = 20.85$$

$$\text{S.D.}(\sigma) = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} \times h = \sqrt{\frac{1190}{200} - \left(\frac{-66}{200}\right)^2} \times 5 = 12.084.$$

NOTES

Calculation of Expected Frequencies

Class	*Lower class limit* x	$z = \frac{x-\mu}{\sigma} = \frac{x-20.85}{12.084}$	*Area under normal curve to the left of ordinate at* z	*Area corresponding to class*	*Expected frequency* $= N \times$ *Area*
0—5	0	– 1.72	0.0427	0.0524	10.48 ≈ **10**
5—10	5	– 1.31	0.0951	0.0890	17.80 ≈ **18**
10—15	10	– 0.90	0.1841	0.1315	26.30 ≈ **26**
15—20	15	– 0.48	0.3156	0.1565	31.30 ≈ **31**
20—25	20	– 0.07	0.4721	0.1610	32.20 ≈ **32**
25—30	25	0.34	0.6331	0.1433	28.66 ≈ **29**
30—35	30	0.76	0.7764	0.1026	20.52 ≈ **21**
35—40	35	1.17	0.8790	0.0639	12.78 ≈ **13**
40—45	40	1.58	0.9429	0.0338	6.76 ≈ **7**
45—50	45	1.99	0.9767	—	—

EXERCISE 10.8

1. Fit a normal curve to the following data:

Variable	60—62	63—65	66—68	69—71	72—74
Frequency	5	18	42	27	8

2. Fit a normal curve to the following data:

Class	60—65	65—70	70—75	75—80
Frequency	3	21	150	335
	80—85	85—90	90—95	95—100
	326	135	26	4

Answers

1. 4, 21, 39, 28, 8 2. 3, 31, 148, 322, 319, 144, 30, 3.

10.36. SUMMARY

- The binomial distribution is a particular type of probability distribution. This was discovered by **James Bernoulli (1654—1705)** in the year 1700. This

NOTES

distribution mainly deals with attributes. An attribute is either present or absent with respect to elements of a population.

- A random variable which counts the number of successes in a random experiment with trials satisfying above four conditions is called a **Binomial variable**.
- The shape of the binomial distribution depends upon the probability of success (p) and the number of trials in the experiment. If $p = q = \frac{1}{2}$, then the distribution will be symmetrical for every value of n. If $p \neq q$, then the distribution would be asymmetrical *i.e.*, skewed. The magnitude of skewness varies as the difference between p and q.
- As number of trials (n) in the binomial distribution increases, the number of successes also increases. If neither p nor q is very small, then as n approaches infinity, the skewness in the distribution disappears and it becomes continuous. We shall see that such a continuous, bell shaped distribution is called a *normal distribution*.
- The Poisson distribution is also a discrete probability distribution. This was discovered by French mathematician **Simon Denis Poisson** (1781 – 1840) in the year 1837. This distribution deals with the evaluation of probabilities of *rare* events such as "no. of car accidents on road", "no. of earthquakes in a year", "no. of misprints in a book", etc.
- The Poisson distribution is derived as a limiting case of binomial distribution.
- A random variable which counts the number of successes in a random experiment with trials satisfying above conditions is called a **Poisson variable.**
- The normal distribution is a particular type of continuous probability distribution. This was discovered by **De Moivre (1667—1754)** in the year 1733. The normal distribution is obtained as a limiting case of a binomial distribution when n, the number of trials is indefinitely large and neither p nor q is very small.

10.37. REVIEW EXERCISES

1. What are the conditions under which Binomial probability model is appropriate
2. Explain the utility of Poisson distribution in practical life.
3. What is a normal probability distribution? What are the salient features of a normal curve?
4. Explain the distinctive features of Binomial and Poisson distributions.
5. What is binomial distribution? Under what conditions will it tend to a normal distribution?
6. What is Poisson distribution? Point out its role.
7. Explain the characteristics of Poisson distribution.
8. Explain the properties of a Binomial distribution. What is its relationship with Poisson distribution?
9. Write short note on Normal distribution.
10. Explain the properties of Normal distribution.
11. How does a normal distribution differ from a binomial distribution? What are the important properties of a normal distribution?
12. Discuss the conditions for the Binomial distribution. What are its important properties?
13. Define binomial distribution and explain its important features.
14. What is meant by theoretical frequency distribution? Discuss the salient features of the Binomial and Normal distributions.
15. Differentiate between Normal and Binomial distributions.
16. What is meant by Theoretical Frequency Distribution? Discuss the main features of Binomial, Poisson and Normal distributions.

11. ESTIMATION THEORY AND HYPOTHESIS TESTING

STRUCTURE

11.1. INTRODUCTION

To describe a set of data or observations, we use statistics such as mean and standard deviation. These statistics are estimated from samples. Sample is nothing but a small section selected from the population and the process of drawing or selecting a sample from the population is called 'sampling'. It is essential that a sample must be a random

NOTES

selection so that each member of the population has the equal chance of being selection in the sample. A statistical population consists of observations of some characteristic of interest associated with the individuals concerned and not the individual items or persons themselves.

A statistical measure based only on all the units selected in a sample is called 'statistic', *e.g.*, sample mean, sample standard deviation, proportion of defectives, etc. whereas a statistical measure based on all the units in the population is called 'parameter'. The terms like mean, median, mode, standard deviation are called parameters when they describe the characteristics of the population and are called statistic when they describe the characteristics of the sample.

A very important aspect of the sampling theory is the study of the tests of significance which enables us to decide on the basis of the sample results whether to accept or reject the hypothesis. A test of significance can be used to compare the characteristics of two samples of the same type. Some of the well known tests of significance for small samples are *t*-test and F-test.

11.2. NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

A statistical hypothesis is a statement about a population parameter. There are two types of statistical hypothesis, null hypothesis and alternative hypothesis.

The hypothesis formulated for the sake of rejecting it under the assumption that it is true, is called the null hypothesis and is denoted by H_0. Null hypothesis asserts that there is no significant difference between the sample statistic and the population parameter and whatever difference is observed that is merely due to fluctuations in sampling from the same population.

Rejecting null hypothesis implies that it is rejected in favour of some other hypothesis which is accepted. A hypothesis which is accepted when H_0 is rejected is called the alternative hypothesis and is denoted by H_1. What we intend to conclude is stated in the alternative hypothesis.

11.3. LEVEL OF SIGNIFICANCE AND CONFIDENCE LIMITS

The probability level below which we reject the hypothesis is known as the 'level of significance'. The region in which a sample value falling is rejected, is known as the 'critical region' or the 'rejection region'. We, generally, take two critical regions which cover 5% and 1% areas of the normal curve.

Depending on the nature of the problem, we use a single-tail test or double-tail test to estimate the significance of a result. In a single-tail test, only the area on the right of an ordinate is taken into consideration whereas in a double-tail test, the areas of both the tails of the curve representing the sampling distribution are taken into consideration.

For example, a test for testing the mean of a population

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis $H_1 : \mu > \mu_0$ (right tailed) or $H_1 : \mu < \mu_0$ (left tailed) is a single tailed test. In the right tailed test ($H_1 : \mu > \mu_0$), the critical region lies entirely

NOTES

in the right tail of the sampling distribution; while for the left tail test ($H_1 : \mu < \mu_0$), the critical region is entirely in the left tail of the sampling distribution.

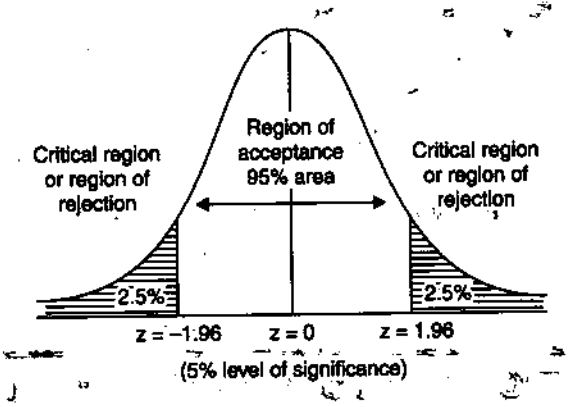
A test of statistical hypothesis where the alternative hypothesis is two tailed such as:

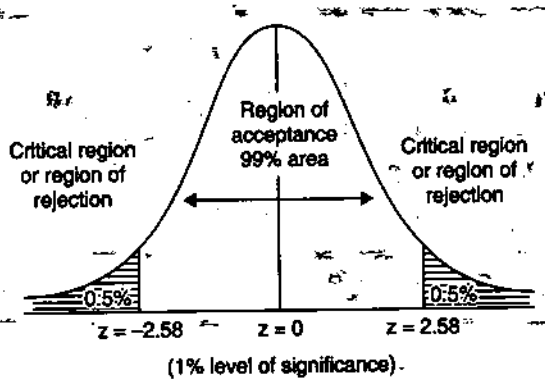
$H_0 : \mu = \mu_0$ against the alternative hypothesis

$H_1 : \mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$) is known as two tailed test and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistic.

The value of z corresponding to 5% level of significance is ± 1.96 and corresponding to 1% level of significance value of z is ± 2.58. The set of z-scores outside the range ± 1.96 and ± 2.58 constitute the critical region of the hypothesis (or the region of rejection) at 5% and 1% level of significance respectively.

The following figure showing region of acceptance and rejection for 5% and 1% level of significance.





11.4. TYPE I ERROR AND TYPE II ERROR

The error of rejecting H_0 when H_0 is true is called the type I error and the error of accepting H_0 when H_0 is false (H_1 is true) is called the type II error. The probability of type I error is denoted by α and the probability of type II error is denoted by β.

P (rejecting H_0 when H_0 is true) = α

P (accepting H_0 when H_1 is true) = β

NOTES

11.5. POWER OF THE TEST

A good test should accept the null hypothesis when it is true and reject the null hypothesis when it is false. $1 - \beta$ (*i.e.*, 1–probability of type II error) measures how well the test is working and is called the power of the test.

Power of the test = $1 - \beta$.

I. TEST OF SIGNIFICANCE FOR SMALL SAMPLES

11.6. STUDENT'S t-TEST

Let $x_1, x_2, \ldots\ldots, x_n$ be a random sample of size n ($n < 30$) from a normal population with mean μ and variance σ^2. The student's t-test is defined as

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is the sample mean and $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is an unbiased estimate of the standard deviation σ.

11.7. ASSUMPTIONS FOR STUDENT'S t-TEST

The following assumptions are made in student's t–test:

(*i*) The parent population from which the sample is drawn is normal.

(*ii*) The population standard deviation (σ) is unknown.

(*iii*) Sample size is less than 30.

11.8. DEGREE OF FREEDOM

The number of independent variates which make up the statistic is known as the degree of freedom (d.f.) and is denoted by ν (the letter 'Nu' of the Greek alphabet).

In general the degree of freedom is defined as

d.f. = number of frequencies – number of independent constraints on them.

11.9. TEST FOR SINGLE MEAN

Suppose we want to test

(*i*) If a random sample x_i ($i = 1, 2, \ldots\ldots, n$) of size n has been drawn from a normal population with a specified mean say μ or

(*ii*) If the sample mean differs significantly from the hypothetical value μ of the population mean.

Under null hypothesis H_0:

(*i*) The sample mean has been drawn from the population with mean μ or

(*ii*) There is no significant difference between the sample mean $\bar{x}$ and the population mean μ, the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

follows Student's t-distribution with $(n-1)$ degrees of freedom.

We now compare the calculated value of t with the tabulated value at certain level of significance. If calculated $|t|$ > tabulated t, H_0 is rejected and if calculated $|t|$ < tabulated t, H_0 may be accepted.

Note. We know, the sample variance

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

$$\therefore \quad ns^2 = (n-1)\,S^2$$

or $$\frac{S^2}{n} = \frac{s^2}{n-1} \Rightarrow \frac{S}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

Hence, the test statistic becomes

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

Example 11.1. *The mean weekly sales of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful?*

Solution. Here, $n = 22$, $\bar{x} = 153.7$, $s = 17.2$

Null hypothesis $H_0 : \mu = 146.3$, *i.e.*, the advertising campaign is not successful.

Alternative hypothesis $H_1 : \mu > 146.3$ (Right tail)

Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \text{ with } (22-1) = 21 \text{ d.f.}$$

$$t = \frac{153.7 - 146.3}{17.2/\sqrt{22-1}} = \frac{7.4 \times \sqrt{21}}{17.2} = 9.$$

Since calculated value of $t = 9$ is greater than the tabulated value of $t = 1.72$ for 21 d.f. at 5% level of significance. It is highly significant. So H_0 is rejected, *i.e.*, the advertising campaign was successful in promoting sales.

Example 11.2. *Ten individuals are chosen at random from a normal population and the heights are found to be in inches 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71. Test if the sample belongs to the population whose mean height is 66 inches. (Given $t_{0.05} = 2.26$ for 9 d.f.)*

NOTES

Solution.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
63	– 4.8	23.04
63	– 4.8	23.04
66	– 1.8	3.24
67	– 0.8	0.64
68	0.2	0.04
69	1.2	1.44
70	2.2	4.84
70	2.2	4.84
71	3.2	10.24
71	3.2	10.24
$\Sigma x_i = 678$		$\Sigma(x_i - \bar{x})^2 = 81.6$

Here, $n = 10$

$$\bar{x} = \text{sample mean} = \frac{\Sigma x_i}{n} = \frac{678}{10} = 67.8 \text{ inches}$$

$$S = \sqrt{\frac{1}{n-1}\Sigma(x_i - \bar{x})^2} = \sqrt{\frac{1}{9} \times 81.6}$$

$$= \sqrt{9.0667} = 3.011$$

Null hypothesis $H_0 : \mu = 66$, *i.e.*, population mean is 66 inches

Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{67.8 - 66}{3.011/\sqrt{10}} = \frac{1.8 \times \sqrt{10}}{3.011}$$

$$= \frac{5.692}{3.011} = 1.8904$$

degree of freedom $= n - 1 = 10 - 1 = 9$

$t_{0.05} = 2.26$ for 9 d.f.

As the calculated value of $|t|$ is less than $t_{0.05}$, the difference between $\bar{x}$ and μ may be due to fluctuations of random sampling. H_0 may be accepted. In other words, the data does not provide any significant evidence against the hypothesis that the population mean is 66 inches.

Example 11.3. *A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43.5 inches for the population is not reasonable. (Given $t_{0.05} = 2.13$, $t_{0.01} = 2.95$ for 15 degrees of freedom)*

Solution. Here, $\bar{x} = 41.5$ inches, $n = 16$, $\Sigma(x_i - \bar{x})^2 = 135$ sq. inches

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{15} \times 135} = \sqrt{9} = 3$$

Null hypothesis $H_0 : \mu = 43.5$ inches, *i.e.*, the data are consistent with an assumption that the mean height in population is 43.5 inches.

Alternative hypothesis $H_1 : \mu \neq 43.5$ inches

Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$|t| = \frac{|41.5 - 43.5|}{3/\sqrt{16}} = \frac{2 \times 4}{3} = 2.667$$

degrees of freedom $= n - 1 = 16 - 1 = 15$

We are given $t_{0.05} = 2.13$ and $t_{0.01} = 2.95$ for 15 degrees of freedom.

Since calculated $|t|$ is greater than $t_{0.05} = 2.13$, null hypothesis H_0 is rejected at 5% level of significance and we conclude that the assumption of mean 43.5 inches for the population is not reasonable.

Remark. Since calculated $|t|$ is less than $t_{0.01} = 2.95$, null hypothesis H_0 may be accepted at 1% level of significance.

11.10. t-TEST FOR DIFFERENCE OF MEANS

Given two independent random samples x_i $(i = 1, 2, \ldots\ldots, n_1)$ and y_j $(j = 1, 2, \ldots\ldots, n_2)$ of sizes n_1 and n_2 with means $\bar{x}$ and $\bar{y}$ and standard deviations S_1 and S_2 from normal populations with the same variance, we have to test the hypothesis that the population means are same. In other words, since a normal distribution is completely specified by its mean and variance, we have to test the hypothesis that the two independent samples come from the same normal population.

The statistic is given by

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i\,;\ \bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j$$

and $$S^2 = \frac{1}{(n_1 + n_2 - 2)}\left[(n_1 - 1)\,S_1^{\,2} + (n_2 - 1)S_2^{\,2}\right]$$

or $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{j=1}^{n_2}(y_j - \bar{y})^2\right]$$

follows Student's t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

If the calculated value of $|t|$ be > tabulated t, the difference between the sample means is said to be significant at certain level of significance; otherwise the data are said to be consistent with the hypothesis.

11.11. PAIRED t-TEST FOR DIFFERENCE OF MEANS

If the size of the two samples is the same, say equal to n, and the data are paired, (x_i, y_i), $(i = 1, 2, \ldots\ldots, n)$ corresponds to the same ith sample unit. The problem is to test if the sample means differ significantly or not.

Here, we consider the increments, $d_i = x_i - y_i$, $(i = 1, 2, \ldots\ldots, n)$.

Under the null hypothesis H_0 that increments are due to fluctuations of sampling, the statistic

$$t = \frac{\bar{d}}{S/\sqrt{n}},$$

where $$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

follows Student's t-distribution with $(n - 1)$ degrees of freedom. If Σd_i is negative, we may consider $|\bar{d}|$. This test is generally one tailed test. Therefore, the alternative hypothesis is $H_1 : \mu_1 > \mu_2$ or $H_1 : \mu_1 < \mu_2$.

NOTES

Example 11.4. *The following data related to the heights (in cms) of two different varieties of wheat plants.*

Variety 1	*63*	*65*	*68*	*69*	*71*	*72*				
Variety 2	*61*	*62*	*65*	*66*	*69*	*69*	*70*	*71*	*72*	*73*

Test the null hypothesis that the mean heights of plants of both varieties are the same.

Solution. Given $n_1 = 6$, $n_2 = 10$

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 > \mu_2$ (right tail)

Under H_0 the test statistic is given by

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Variety 1

x	$x - \bar{x} = x - 68$	$(x - \bar{x})^2$
63	– 5	25
65	– 3	9
68	0	0
69	1	1
71	3	9
72	4	16
$\Sigma x = 408$		$\Sigma(x - \bar{x})^2 = 60$

Variety 2

y	$y - \bar{y} = y - 67$	$(y - \bar{y})^2$
61	– 6	36
62	– 5	25
65	– 2	4
65	– 2	4
66	– 1	1
66	– 1	1
70	3	9
70	3	9
72	5	25
73	6	36
$\Sigma y_i = 670$		$\Sigma(y - \bar{y})^2 = 150$

$$\bar{x} = \frac{1}{n_1}\Sigma x_i = \frac{408}{6} = 68 \qquad \bar{y} = \frac{1}{n_2}\Sigma y_i = \frac{670}{10} = 67$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2]$$

$$= \frac{1}{6 + 10 - 2}[60 + 150] = \frac{210}{14} = 15 \Rightarrow S = 3.873$$

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{68 - 67}{3.873\sqrt{\frac{1}{6} + \frac{1}{10}}} = \frac{1}{3.873 \times 0.5164} = 0.499$$

Tabulated $t_{0.05}$ for 14 degrees of freedom for single tail-test is 1.76.

Since calculated value of t is less than 1.76, it is not at all significant at 5% level of significance. Hence, H_0 may be accepted and we conclude that the height of the plants are not different at 5% level of significance.

Example 11.5. *The mean values of birth weight with standard deviations and sample sizes are given below by socio-economic status. Is the mean difference in birth weight significant between socio-economic group?*

	High socio-economic graph	*Low socio-economic group*
Sample size	$n_1 = 15$	$n_2 = 10$
Birth weight (kg)	$\bar{x} = 2.91$	$\bar{y} = 2.26$
Standard deviation	$S_1 = 0.27$	$S_2 = 0.22$

Solution. Given $n_1 = 15$, $n_2 = 10$, $\bar{x} = 2.91$, $\bar{y} = 2.26$

$S_1 = 0.27$ and $S_2 = 0.22$

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 > \mu_2$ (right tail), *i.e.* high socio-economic group is superior to low socio-economic group.

Under H_0 the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\,S_1^2 + (n_2 - 1)\,S_2^2\right]$$

$$= \frac{1}{15 + 10 - 2}\left[(15 - 1) \times (0.27)^2 + (10 - 1) \times (0.22)^2\right]$$

$$= \frac{1.0206 + 0.4356}{23} = \frac{1.4562}{23} = 0.063$$

$$\Rightarrow \quad S = 0.25$$

$$\therefore \quad t = \frac{2.91 - 2.26}{0.25\sqrt{\frac{1}{15} + \frac{1}{10}}} = \frac{0.65 \times \sqrt{150}}{0.25 \times \sqrt{25}} = \frac{0.65 \times 2.45}{0.25} = 6.37$$

Tabulated value of t for 23 degrees of freedom at 5% level of sign. Finance for right tailed test is 1.71. Since calculated t is much greater than tabulated t, it is highly significance and H_0 is rejected and conclude that mean of high group is greater than low group.

Example 11.6. *Memory capacity of 8 students was tested before and after training. State at 5% level of significance whether the training was effective from the following scores:*

Student	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Total*
Before	*49*	*53*	*51*	*52*	*47*	*50*	*52*	*53*	*407*
After	*52*	*55*	*52*	*53*	*50*	*54*	*54*	*53*	*423*

Use paired t-test for your answer.

Solution. Let x denotes the scores before training and y denotes the scores after training.

Null hypothesis $H_0 : \mu_1 = \mu_2$, *i.e.* there is no significant difference in the scores before and after the training. In other words, the given increments are just by chance (fluctuations of sampling).

NOTES

Alternative hypothesis $H_1 : \mu_1 < \mu_2$ (to conclude that training has been effected) (One tail)

Student	*Score before training (x)*	*Score after training (y)*	$d = x - y$	d^2
1	49	52	– 3	9
2	53	55	– 2	4
3	51	52	– 1	1
4	52	53	– 1	1
5	47	50	– 3	9
6	50	54	– 4	16
7	52	54	– 2	4
8	53	53	0	0
			$\Sigma d = -16$	$\Sigma d^2 = 44$

Under H_0 the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}}$$

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{-16}{8} = -2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{n-1}[\Sigma d_i^2 - n(\bar{d})^2]$$

$$= \frac{1}{7}[44 - 8 \times (-2)^2] = \frac{44-32}{7} = \frac{12}{7} = 1.714$$

$$\Rightarrow \quad S = 1.31$$

$$\therefore \quad |t| = \frac{|d|}{S/\sqrt{n}} = \frac{|-2|}{1.31/\sqrt{8}} = \frac{2 \times 2.83}{1.31} = 4.32$$

Tabulated $t_{0.05}$ for $(8 - 1) = 7$ degrees of freedom for one tail test is 1.90.

Since calculated value of t is greater than the tabulated t, H_0 is rejected at 5% level of significance. Hence, we conclude that the scores differ significantly before and after the training, *i.e.* training was effected.

EXERCISE 11.1

1. A brand of matches is sold in boxes on which it is claimed that the average contents are 40 matches. A check on a pack of 5 boxes gives the following results:

   41, 39, 37, 40, 38

   (*i*) Test the manufacturer's claim keeping the interests of both the manufacturer and the customer in mind.

   (*ii*) As a customer test the manufacturer's claim.

2. A sample of size 10 drawn from a normal population has a mean 31 and a variance 2.25. Is it reasonable to assume that the mean of the population is 30? (Use 1% level of significance).

3. A random sample of size 10 from a normal population with mean μ gives a sample mean of 40 and sample standard deviation of 6. Test the hypothesis that $\mu = 44$ against $\mu \neq 44$ at 5% level of significance.

4. A new drug manufacturer wants to market a new drug only if he could be quite sure that the mean temperature of a healthy person taking the drug could not rise above 98.6°F otherwise he will withhold the drug. The drug is administered to a random sample of 17 healthy persons. The mean temperature was found to be 98.4°F with a standard deviation of 0.6°F. Assuming that the distribution of the temperature is normal and $\alpha = 0.01$, what should the manufacturer do?

5. The marks of students in two groups were obtained as

I	18	20	36	50	49	36	34	49	41
II	29	28	26	35	30	44	46		

Test whether the groups were identical.
(Given $t_{0.05} = 2.14$ for 14 degrees of freedom)

6. Two different types of drugs A and B were tried on certain patients for increasing weight. 5 persons were given drug A and 7 persons were given drug B. The increase in weight in pounds is given below:

Drug A	8	12	13	9	3		
Drug B	10	8	12	15	6	8	11

Do the two drugs differ significantly with regard to their effect in increasing weight.
(Given $t_{0.05} = 2.23$ for 10 degrees of freedom)

7. The mean life of a sample of 10 electric light bulbs was found to be 1456 hours with standard deviation of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with standard deviation of 398 hours. Is there a significant difference between the means of the two batches?
(Given $t_{0.05} = 2.06$ for 25 degrees of freedom)

8. To verify whether a course in Statistics improved performance, a similar test was given to 12 participants both before and after the course. The original marks recorded in alphabetical order of the participants were 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. After the course, the marks were in the same order 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78. Was the course useful?
(Given $t_{0.05} = 2.201$ for 11 degrees of freedom)

9. A certain medicine given to each of the 9 patients resulted in the following increase of blood pressure. Can it be concluded that the medicine will in general be accompanied by an increase in blood pressure.
7, 3, – 1, 4, – 3, 5, 6, – 4, – 1
(Given $t_{0.05} = 2.306$ for 8 degrees of freedom)

Answers

1. (i) Accept manufacturer 's claim (ii) manufacturer's claim is justified.
2. Yes
3. Accept null hypothesis
4. The manufacturer should market the drug
5. Two groups are identical
6. No
7. No
8. Yes
9. No

11.12. F-TEST

This test uses the variance ratio to test the significance of difference between two sampled variances. F-test which is based on F-distribution is called so in honour of a great statistician Prof. R.A. Fisher.

NOTES

Let $x_1, x_2, \ldots\ldots, x_{n_1}$ and $y_1, y_2, \ldots\ldots, y_{n_2}$ be the values of two independent random samples drawn from the same normal population with variance σ^2. Then, we define variance ratio F as follows:

$$F = \frac{S_1^2}{S_2^2} ; S_1 > S_2$$

where
$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2$$

and $\bar{x}, \bar{y}$ are the sample means.

The distribution of variance ratio F with ν_1 and ν_2 degrees of freedom is given by

$$y = \frac{y_0 F^{\left(\frac{\nu_1 - 2}{2}\right)}}{\left(1 + \frac{\nu_1}{\nu_2} F\right)^{\left(\frac{\nu_1 + \nu_2}{2}\right)}},$$

where y_0 is so chosen that the total area under the curve is unity.

The parameters ν_1 and ν_2 represent degrees of freedom. For samples of sizes n_1 and n_2, we have

$$\nu_1 = n_1 - 1 \quad \text{and} \quad \nu_2 = n_2 - 1.$$

11.13. PROPERTIES OF F-DISTRIBUTION

(*i*) The value of F cannot be negative as both terms of F-ratio are the squared values.

(*ii*) The range of the values of F is from 0 to ∞.

(*iii*) The F-distribution is independent of the population variance σ^2 and depends on ν_1 and ν_2 only.

The F-distribution for various degrees of freedom ν_1 and ν_2 is given in the following table:

Table: Values of F for 5% and 1% level, where ν_1 is the number of degree of freedom for greater estimate of variance and ν_2 for the smaller estimate of variance.

11.14. PROCEDURE TO F-TEST

(*i*) Set up the null hypothesis $H_0 = \sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.* the independent estimates of the common population variance do not differ significantly.

(*ii*) Find the degrees of freedom ν_1 and ν_2 given by $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ respectively.

(*iii*) Calculate the variances of two samples and then calculate F.

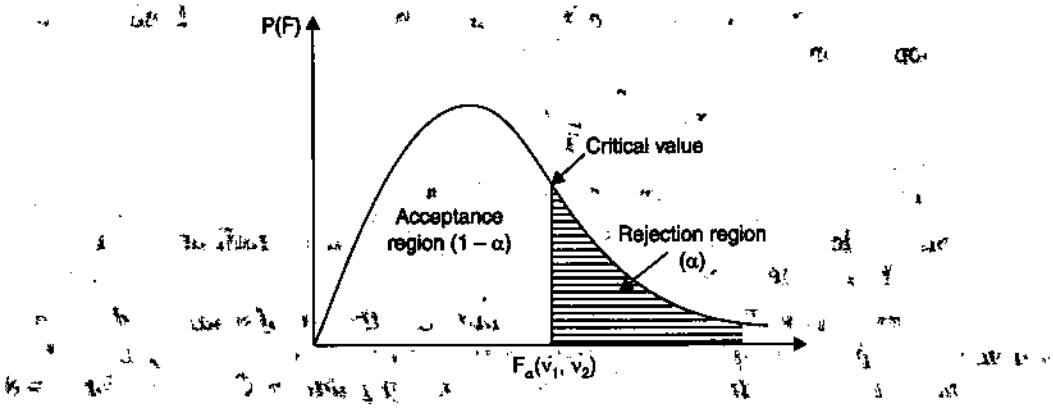
(*iv*) From F-distribution table note the value of F for ν_1, ν_2 degrees of freedom at the desired level of significance.

(*v*) Compare the calculated value of F with tabulated value of F at the desired level of significance. If the calculated value of F is less than the tabulated value, then the difference is not significant and we may conclude that the same could have come from two populations with the same variance *i.e.*, accept H_0, otherwise reject H_0.

11.15. CRITICAL VALUES OF F-DISTRIBUTION

The available F-table give the critical values of F for the right-tailed test, *i.e.* the critical region is determined by the right-tail areas. Thus, the significance value $F_\alpha(\nu_1, \nu_2)$ at level of significance and (ν_1, ν_2) degrees of freedom is determined by

$P[F > F_\alpha(\nu_1, \nu_2)] = \alpha$, as shown below:



Example 11.7. *In one sample of size 8 the sum of the squares of deviations of the sample values from the sample mean is 84.4 and in the other sample of size 10 it is 102.6. Test whether this difference is significance at 5% level. Given that for $\nu_1 = 7$ and $\nu_2 = 9$; $F_{0.05} = 3.29$.*

Solution. Here, $n_1 = 8$, $n_2 = 10$

and $\Sigma(x - \bar{x})^2 = 84.4$, $\Sigma(y - \bar{y})^2 = 102.6$

$$S_1^2 = \frac{1}{n_1 - 1}\Sigma(x - \bar{x})^2 = \frac{1}{7} \times 84.4 = 12.057$$

$$S_2^2 = \frac{1}{n_2 - 1}\Sigma(y - \bar{y})^2 = \frac{1}{9} \times 102.6 = 11.4$$

Under $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.* the estimates of σ^2 given by the samples are homogeneous,

$$F = \frac{S_1^2}{S_2^2} = \frac{12.057}{11.4} = 1.057$$

For $\nu_1 = 7$ and $\nu_2 = 9$, we have $F_{0.05} = 3.29$. Since calculated value of F is less than $F_{0.05}$, H_0 may be accepted at 5% level of significance.

Example 11.8. *Two random samples gave the following information:*

Sample	*Size*	*Sample mean*	*Sum of squares of deviations from the mean*
1	*10*	*15*	*90*
2	*12*	*14*	*108*

Test whether the samples have been drawn from the same normal population. Given that for $\nu_1 = 9$ and $\nu_2 = 11$; $F_{0.05} = 2.90$ (approx.).

NOTES

Solution. Here, $n_1 = 10$, $n_2 = 12$, $\bar{x} = 15$, $\bar{y} = 14$

$$\Sigma (x - \bar{x})^2 = 90; \Sigma(y - \bar{y})^2 = 108$$

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{1}{9} \times 90 = 10$$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1}{11} \times 108 = 9.82$$

Under $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.* two samples have been drawn from the same normal population.

$$F = \frac{S_1^2}{S_2^2} = \frac{10}{9.82} = 1.018$$

For $\nu_1 = 9$ and $\nu_2 = 11$, we have $F_{0.05} = 2.90$.

Since calculated value of F is less than $F_{0.05}$ it is not significant. Hence, null hypothesis H_0 may be accepted.

Example 11.9. *The samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 and 91 square units respectively. Test whether the samples have been drawn from the same normal population. Given that for $\nu_1 = 8$ and $\nu_2 = 7$; $F_{0.05} = 3.73$.*

Solution. Here, $n_1 = 9$, $n_2 = 8$, $\Sigma(x - \bar{x})^2 = 160$, $\Sigma(y - \bar{y})^2 = 91$

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{1}{8} \times 160 = 20$$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1}{7} \times 91 = 13$$

Under $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.* two samples have been drawn from the same normal population.

$$F = \frac{S_1^2}{S_2^2} = \frac{20}{13} = 1.54 \text{ (approx.)}$$

For $\nu_1 = 8$ and $\nu_2 = 7$, we have $F_{0.05} = 3.73$

Since calculated value of F is less than $F_{0.05}$ it is not significant. Hence, H_0 may be accepted.

Example 11.10. *Two samples are drawn from two normal populations. From the following data test whether the two samples have the same variances at 5% level of significance.*

Sample I	*60*	*65*	*71*	*74*	*76*	*82*	*85*	*87*		
Sample II	*61*	*66*	*67*	*85*	*78*	*88*	*86*	*85*	*63*	*91*

Solution. Here, $n_1 = 8$, $n_2 = 10$

Under $H_0 : S_1^2 = S_2^2$, *i.e.* two samples have the same variance.

$$H_1 : S_1^2 \neq S_2^2$$

Sample-I			Sample-II		
x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-\bar{y})^2$
60	60–75 = – 15	225	61	61–77 = – 16	256
65	65–75 = – 10	100	66	66–77 = – 11	121
71	71–75 = – 4	16	67	67–77 = – 10	100
74	74–75 = – 1	1	85	85–77 = 8	64
76	76–75 = 1	1	78	78–77 = 1	1
82	82–75 = 7	49	88	88–77 = 11	121
85	85–75 = 10	100	86	86–77 = 9	81
87	87–75 = 12	144	85	85–77 = 8	64
			63	63–77 = – 14	196
			91	91–77 = 14	196
$\Sigma x = 600$		$\Sigma(x-\bar{x})^2$ = 636	$\Sigma y = 770$		$\Sigma(y-\bar{y})^2$ = 1200

$$\bar{x} = \frac{\Sigma x}{n_1} = \frac{600}{8} = 75 \qquad \bar{y} = \frac{\Sigma y}{n_2} = \frac{770}{10} = 77$$

$$\text{Variance of sample-I} = S_1^2 = \frac{1}{n_1 - 1}\Sigma(x-\bar{x})^2 = \frac{636}{8-1} = 90.857$$

$$\text{Variance of sample-II} = S_2^2 = \frac{1}{n_2 - 1}\Sigma(y-\bar{y})^2 = \frac{1200}{10-1} = 133.33$$

$$F = \frac{S_2^2}{S_1^2} = \frac{133.33}{90.857} = 1.467$$

For $\nu_1 = 7$ and $\nu_2 = 9$, we have $F_{0.05} = 3.29$.

Since calculated value of F is less than $F_{0.05}$, H_0 may be accepted, *i.e.* the samples I and II have the same variance.

EXERCISE 11.2

1. In a sample of 8 observations, the sum of squared deviations of items from the mean was 94.5. In another sample of 10 observations, the value was found to be 101.7. Test whether the difference is significant at 5% level.
2. The following are the values in thousands of an inch obtained by two engineers in 10 successive measurements with the same micrometer. Is one engineer significantly more consistent than the other?

Engineer A	503	505	497	505	495	502	499	493	510	501
Engineer B	502	497	492	498	499	495	497	496	498	

3. The nicotine content (in milligrams) of two samples of tobacco were found to be as follows:

Sample A	24	27	26	21	25	
Sample B	27	30	28	31	22	36

Can it be said that the two samples come from the same normal population?

NOTES

4. The daily wages in ₹ of skilled workers in two cities are as follows:

City	*Size of sample of workers*	*S.D. of wages in the sample*
A	16	25
B	13	32

Test at 5% level of significance the equality of variances of the wage distribution in the two cities.

5. The time taken by workers in performing a job by methods I and II is given below:

Method I	20	16	26	27	23	22	–
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution from population from which these samples are drawn do not differ significantly?

6. Two random samples drawn from two normal populations are given below:

Sample I	63	65	68	69	71	72	–	–	–	–
Sample II	63	62	65	66	69	69	70	71	72	73

Test whether the two populations have the same variance at 5% level of significance.

Answers

1. No
2. Not significant
3. yes
4. Accepted
5. Not significant
6. Yes.

II. TEST OF SIGNIFICANCE FOR LARGE SAMPLES

For practical purposes a sample is taken as a large sample if $n > 30$. Under large sample test there are some important tests to test the significance. These tests are as follows:

1. Test of significance for proportion
   (*i*) Single proportion (*ii*) Difference of proportions
2. Test of significance for single mean.
3. Test of significance for differences of
   (*i*) Means (*ii*) Standard deviations.

11.16. TEST OF SIGNIFICANCE FOR PROPORTION

(*i*) Single proportion: This test is used to test the significant difference between proportion of the sample and the population.

Let X be the number of successes in n independent trials with constant probability P of success for each trial.

We have $E(X) = nP$ and $V(X) = nPQ$, where $Q = 1 - P$ = probability of failure

Now, $p = \frac{X}{n}$ (p = observed proportion of success)

Now, $$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{nP}{n} = P$$

$$V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{nPQ}{n^2} = \frac{PQ}{n}$$

$$S.E.\ (p) = \sqrt{\frac{PQ}{n}}$$

$$Z = \frac{p - E(p)}{S.E.\ (p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0, 1)$$

where E → expected value, V → Variance and S.E. → Standard error

Z is called a test statistic which is used to test the significant difference of the sample and population proportion.

Note 1. The probable limits for the observed proportion of success are $E(p) \pm Z_\alpha \sqrt{V(p)}$ *i.e.*, $P \pm Z_\alpha \sqrt{\frac{PQ}{n}}$, where Z_α is the significant value at the level of significance α.

2. If P is not known then the probable limits for the proportion in the population are $p \pm Z_\alpha \sqrt{\frac{pq}{n}}$.

3. If α is not given, then we can use 3σ limits. Hence, probable limits for the observed proportion of success are $P \pm 3\sqrt{\frac{PQ}{n}}$ and probable limits for the proportion in the population are $p \pm 3\sqrt{\frac{pq}{n}}$.

4. A set of four selected values is commonly used for α. Each α and corresponding Z_α and $Z_{\alpha/2}$ values are given in the following table:

For two-tailed test		*For one-tailed test*	
α	$Z_{\alpha/2}$	α	Z_α
0.20	1.282	0.10	1.282
0.10	1.645	0.05	1.645
0.05	1.960	0.025	1.960
0.01	2.576	0.01	2.326

***(ii)* Difference of Proportions:** This test is used to test the difference between the sample proportions.

Let two samples X_1 and X_2 of sizes n_1 and n_2 respectively taken from two different populations, then $p_1 = \frac{X_1}{n_1}$ and $p_2 = \frac{X_2}{n_2}$.

To test the significance of the difference between the sample proportions p_1 and p_2 we set the null hypothesis H_0, that there is no significant difference between the two sample proportion.

NOTES

Under the null hypothesis H_0, the test statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \text{ and } Q = 1 - P.$$

If sample proportions are not given, we set the null hypothesis

$$H_0 : p_1 = p_1$$

under H_0 the test statistic is

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}, \text{ where } Q_1 = 1 - P_1 \text{ and } Q_2 = 1 - P_2.$$

Example 9.11. *A coin is tossed 324 times and the head turned up 175 times. Test the hypothesis that the coin is unbiased.*

Solution. Null hypothesis H_0 : the coin is unbiased *i.e.*,

$$P = \frac{1}{2}$$

Here, $n = 324$, X = Number of heads = 175

P = prob. of getting a head in a toss = $\frac{1}{2}$

$$Q = 1 - P = 1 - \frac{1}{2} = \frac{1}{2}$$

$$Z = \frac{X - E(X)}{\text{SE of } X} = \frac{X - nP}{\sqrt{nPQ}} = \frac{175 - 324 \times \frac{1}{2}}{\sqrt{324 \times \frac{1}{2} \times \frac{1}{2}}}$$

$$= \frac{13}{9} = 1.44 < 1.96$$

Since | Z | < 1.96, null hypothesis is accepted at 5% level of significance. Hence the coin is unbiased.

Example 9.12. *A die is thrown 1000 times and a throw of 5 or 6 was obtained 420 times. On the assumption of random throwing do the data indicate an unbiased die?*

Solution. Null hypothesis H_0 : the die is unbiased

Under H_0, P = probability of getting 5 or 6

$$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

$$Q = 1 - P = 1 - \frac{1}{3} = \frac{2}{3}$$

Here, $n = 1000$, X = Number of success = 420

$$Z = \frac{X - nP}{\sqrt{nPQ}} = \frac{420 - 1000 \times \frac{1}{3}}{\sqrt{1000 \times \frac{1}{3} \times \frac{2}{3}}} = \frac{420 - 333.33}{\sqrt{222.222}} = \frac{86.67}{14.91} = 5.813$$

Since | Z | = 5.813 > 3 (Maximum value of Z), H_0 is rejected *i.e.*, the die is biased.

Example 11.13. *500 apples are taken at random from a large basket and 65 are found to be bad. Find the S.E. of the proportion of bad ones in a sample of this size and assign limits within which the percentage of bad apples most probably lies.*

Solution. Here, $n = 500$, X = number of bad apples in the sample = 65

p = proportion of bad apples in the sample = $\frac{65}{500} = 0.13$ and

$q = 1 - p = 1 - 0.13 = 0.87$

∵ The proportion of bad apples P in the population is not known.

∴ We can take $P = p = 0.13$, $Q = q = 0.87$ and $N = n = 500$

S.E. of proportion $= \sqrt{\frac{PQ}{N}} = \sqrt{\frac{0.13 \times 0.87}{500}} = 0.015$

Limits for proportions of bad apples in the population is

$$P \pm 3\sqrt{\frac{PQ}{N}} = 0.13 \pm 3\sqrt{\frac{0.13 \times 0.87}{500}} = 0.13 \pm 0.045 = 0.175 \text{ and } 0.085$$

$$= 17.5\% \text{ and } 8.5\%.$$

Example 11.14. *Before an increase in excise duty on tea, 400 people out of a sample of 500 persons where found to be tea drinkers. After an increase in the excise duty, 400 persons were known to be tea drinkers in a sample of 600 people. Do you think that there has been a significant decrease in the consumption of tea after the increase in the excise duty?*

Solution. Here $n_1 = 500$, $n_2 = 600$

$X_1 = 400$, $X_2 = 400$

p_1 = proportion of drinkers in first sample $= \frac{400}{500} = \frac{4}{5} = 0.8$

p_2 = proportion of drinkers in second sample $= \frac{400}{600} = \frac{2}{3} = 0.67$

Since proportion P of the population is not given, it can be estimated by using

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400 + 400}{500 + 600} = \frac{800}{1100} = \frac{8}{11}$$

and $Q = 1 - P = 1 - \frac{8}{11} = \frac{3}{11}$

Null hypothesis $H_0 : P_1 = P_2$ (there is no significant difference in the consumption of tea before and after increase of excise duty)

Alternative hypothesis $H_1 : P_1 > P_2$ (right tailed test), under H_0 the test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.8 - 0.67}{\sqrt{\frac{8}{11} \times \frac{3}{11}\left(\frac{1}{500} + \frac{1}{600}\right)}} = \frac{0.13}{0.027} = 4.815$$

Since $|Z| = 4.815 > 1.645$ also $|Z| = 4.815 > 2.33$ at both the significant values of Z at 5% and 1% level of significant respectively, H_0 is rejected *i.e.*, there is a significant decrease in the consumption of tea due to increase in excise duty.

NOTES

Example 11.15. *500 articles from a factory are examined and found to be 2% defective. 800 similar articles from a second factory are found to have only 1.5% defectives. Can it reasonably concluded that the products of the first factory are inferior to those of second?*

Solution. Here, $n_1 = 500$,

p_1 = proportion of defectives from first factory = $\frac{2}{100} = 0.02$

$n_2 = 800$,

p_2 = proportion of defectives from second factory = $\frac{15}{100} = 0.015$

Since proportion P of the population is not given it can be estimated by using

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{10 + 12}{500 + 800} = \frac{22}{1300} = 0.017$$

and $$Q = 1 - P = 1 - 0.017 = 0.983$$

Null hypothesis $H_0 : P_1 = P_2$ (there is no significant difference between the products of first and second factory)

Alternative hypothesis $H_1 : P_1 \neq P_2$ (two tailed test)

Under H_0 the test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.02 - 0.015}{\sqrt{0.017 \times 0.983\left(\frac{1}{500} + \frac{1}{800}\right)}}$$

$$= \frac{0.005}{0.00737} = 0.678$$

Since $|Z| = 0.678 < 1.96$, null hypothesis is accepted at 5% level of significance. Hence there is no significant difference between the products of first and second factory *i.e.*, the products of the first factory are not inferior to those of second.

Example 11.16. *In two large populations there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1400 and 1000 respectively from the two populations.*

Solution. Here, $n_1 = 1400$, $n_2 = 1000$

P_1 = proportion of fair haired in the first population = $\frac{30}{100} = 0.3$

P_2 = proportion of fair haired in the second population = $\frac{25}{100} = 0.25$

$Q_1 = 1 - P_1 = 1 - 0.3 = 0.7$, $Q_2 = 1 - P_2 = 1 - 0.25 = 0.75$

Null hypothesis $H_0 : p_1 = p_2$ (Sample proportions are equal) *i.e.*, the difference in population proportions is likely to be hidden in sampling.

Alternative hypothesis $H_1 : p_1 \neq p_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} = \frac{0.30 - 0.25}{\sqrt{\frac{0.3 \times 0.7}{1400} + \frac{0.25 \times 0.75}{1000}}} = \frac{0.05}{0.01837} = 2.72$$

Since $|Z| = 2.72 > 1.96$, null hypothesis is rejected at 5% level of significance. Hence at 5% level of significance these samples will exhibit the difference in the population proportions.

NOTES

EXERCISE 11.3

1. A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased.
2. In a hospital 525 female and 475 male babies were born in a month. Do these figures confirm the hypothesis that females and males are born in equal number?
3. A die is thrown 10000 times and a throw of 3 or 4 was obtained 4200 times. On the assumption of random throwing do the data indicate an unbiased die?
4. Given that on the average 4% of insured men of age 65 die within a year and that 60 of a particular group of 1000 such men (age 65) died within a year. Can this group be regarded as a representative sample?
5. 325 men out of 600 men chosen from a big city were found to be smokers. Does this information support the conclusion that the majority of men in the city are smokers?
6. A random sample of 400 apples is taken from a large basket and 40 are found to be bad. Estimate the proportion of bad apples in the basket and assign limits within which the percentage most probably lies.
7. A manufacturer claimed that at least 95% of the equipments which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipments revealed that 18 were faulty. Test the manufacturer's claim at a level of significance (*i*) 5% (*ii*) 1%.
8. 1000 articles from a factory are examined and found to be 2.5% defective. 1500 similar articles from a second factory are found to have only 2% defectives. Can it reasonably concluded that the products of the first factory are inferior to those of second?
9. A manufacturing firm claims that its brand A product outsells its brand B product by 8%. If it is found that 42 out of a sample of 200 persons prefer brand A and 18 out of another sample of 100 persons prefer brand B. Test whether the 8% difference is valid claim.
10. In a survey on a particular matter in a college, 850 males and 560 females voted. 500 males and 320 females voted yes. Does this indicate a significant difference of opinion between male and female on this matter at 1% level of significance?
11. Two samples of sizes 1200 and 900 respectively drawn from two large populations. In the two large populations there are 30% and 25% respectively of fair haired people. Test whether these two samples will reveal the difference in the population proportions.
12. Before an increase in excise duty on tea 800 persons out of a sample of 1000 persons were found to be tea drinkers. After an increase in excise duty 800 people were tea drinkers in a sample of 1200 people. Test whether there is a significant decrease in the consumption of tea after the increase in excise duty.

Answers

1. H_0 is accepted at 5% level of significance.
2. Yes, H_0 is accepted at 5% level of significance.
3. H_0 is rejected.
4. H_0 is rejected.
5. H_0 is rejected at 5% level of significance.
6. 8.5 : 11.5
7. Using left tailed test, H_0 is rejected at both 5% and 1% level of significance.
8. No, H_0 is accepted.
9. H_0 is accepted.
10. H_0 is accepted.
11. H_0 is rejected at 5% level of significance.
12. H_0 is rejected.

11.17. TEST OF SIGNIFICANCE FOR SINGLE MEAN

This test is used to test the significant difference between sample mean and population mean.

NOTES

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a normal population with mean μ and variance σ^2.

The standard error (S.E.) of mean of a random sample of size n from a population is given by

S.E. $(\bar{x}) = \frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the population.

We set the null hypothesis H_0 that the sample has been drawn from a large population with mean μ and variance σ^2 *i.e.*, there is no significant difference between the sample mean $(\bar{x})$ and population mean (μ).

Under the null hypothesis H_0 the test statistic is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

If standard deviation of the population (σ) is not known, we use the test statistic given as,

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$, where s is the standard deviation of the sample.

Note. The limits of the population mean μ are given by $\bar{x} \pm Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$ *i.e.*,

$$\bar{x} - Z_\alpha \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

These limits are called the confidence limits for μ.

Example 11.17. *A normal population has a mean of 6.8 and standard deviation of 1.5. A sample of 400 members gave a mean of 6.75. Is the difference significant?*

Solution. Here, $\mu = 6.8$, $\bar{x} = 6.75$, $\sigma = 1.5$, $n = 400$

Null hypothesis H_0 : $\bar{x} = \mu$ (there is no significant difference between $\bar{x}$ and μ)

Alternative hypothesis H_1 : there is a significant difference between $\bar{x}$ and μ

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{6.75 - 6.8}{1.5/\sqrt{400}} = -\frac{0.05}{0.075} = -0.67$$

Since $|Z| = 0.67 < 1.96$ H_0 is accepted at 5% level of significance. Hence there is no significant difference between $\bar{x}$ and μ.

Example 11.18. *A random sample of 400 members has a mean 99. Can it be reasonably regarded as a sample from a large population of mean 100 and standard deviation 8 at 5% level of significance?*

Solution. Here, $\mu = 100$, $\bar{x} = 99$, $\sigma = 8$, $n = 400$

Null hypothesis H_0 : the sample is drawn from a large population with mean 100 and standard deviation 8.

Alternative hypothesis H_1 : $\mu \neq 100$ (two tailed test)

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{99 - 100}{8/\sqrt{400}} = -\frac{1}{0.4} = -2.5$$

Since $|Z| = 2.5 > 1.96$, H_0 is rejected at 5% level of significance. Hence there is a significant difference between $\bar{x}$ and μ *i.e.*, it can not be regarded as a sample from a large population.

Example 11.19. *The management of a company claims that the average weekly income of their employees is ₹ 900. The trade union disputes this claim stressing that it is rather less. An independent sample of 150 randomly selected employees estimated the average to be ₹ 856 with standard deviation of ₹ 354. Would you accept the view of the management?*

Solution. Here, $\mu = 900$, $\bar{x} = 854$, $s = 354$, $n = 150$

Null hypothesis H_0 : there is no significant difference between $\bar{x}$ and μ *i.e.*, the view of management is correct.

Alternative hypothesis $H_1 : \mu \neq 900$ (two-tailed test)

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{854 - 900}{354/\sqrt{150}} = -\frac{46}{28.904} = -1.59$$

Since $|Z| = 1.59 < 1.96$, H_0 is accepted at 5% level of significance. Hence the view of management is correct.

Example 11.20. *In a population with a standard deviation of 14.8, what sample size is needed to estimate the mean of population within ± 1.2 with 95% confidence?*

Solution. Here, $\bar{x} - \mu = \pm 1.2$, $\sigma = 14.8$, $Z = 1.96$

We know that $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Using this, we have

$$1.96 = \frac{\pm 1.2}{14.8/\sqrt{n}} = \frac{\pm 1.2\sqrt{n}}{14.8}$$

On squaring both the sides we have

$$(1.96)^2 = \left(\frac{\pm 1.2}{14.8}\right)^2 \times n \quad \text{or} \quad n = \left(\frac{1.96 \times 14.8}{\pm 1.2}\right)^2 = 584.35 \approx 584.$$

Example 11.21. *A random sample of 900 measurements from a large population gave a mean value of 64. If this sample has been drawn from a normal population with standard deviation of 20, find the 95% and 99% confidence limits for the mean in the population.*

Solution. Here, $n = 900$, $\bar{x} = 64$, $\sigma = 20$

At 95% confidence $Z = 1.96$

At 99% confidence $Z = 2.58$

The confidence limits for the population mean μ is given by

$$\bar{x} \pm Z_\alpha \frac{\sigma}{\sqrt{n}}$$

The confidence limits for 95% confidence are

$$64 \pm 1.96 \times \frac{20}{\sqrt{900}} = 64 \pm 1.307 = 62.693 \text{ and } 65.307$$

The confidence limits for 99% confidence are

$$64 \pm 2.58 \times \frac{20}{\sqrt{900}} = 64 \pm 1.72 = 62.28 \text{ and } 65.72.$$

NOTES

EXERCISE 11.4

1. A random sample of 900 members has a mean 3.4 cms. Can it be reasonably regarded as a sample from a large population of mean 3.2 cms and standard deviation 2.3 cms?
2. A random sample of 400 male students is found to have a mean height of 160 cms. Can it be reasonably regarded as a sample from a large population with mean height 162.5 cms and standard deviation 4.5 cms?
3. A random sample of 200 measurements from a large population gave a mean value of 50 and a standard deviation of 9. Determine 95% confidence interval for the mean of population.
4. A random sample of 400 measurements from a large population gave a mean value of 82 and a standard deviation of 18. Determine 95% confidence interval for the mean of population.
5. A company manufacturing electric bulbs claims that the average life of its bulbs is 1600 hours. The average life and standard deviation of random sample of 100 such bulbs were 1570 hours and 120 hours respectively. Should we accept the claim of the company?
6. An insurance agent has claimed that the average age of policy holders who insure through him is less than the average for all agents which is 30.5 years. A random sample of 100 policy holders who had insured through him reveal that the mean and standard deviation are 28.8 years and 6.35 years respectively. Test his claim at 5% level of significance.
7. The guaranteed average life of a certain type of bulbs is 1000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90% of the bulbs do not fall short of the guaranteed average by more than 2.5%. What must be the minimum size of the sample?

Answers

1. Yes, H_0 is accepted.
2. Yes, H_0 is accepted.
3. 48.8 and 51.2
4. 80.24 and 83.76
5. No, rejected at 5% level of significance.
6. Claim is valid.
7. $n = 4$

11.18. TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS

(*i*) This test is used to test the significant difference between the means of two large samples.

Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let $\bar{x}_2$ be the mean of an independent sample of size n_2 from another population with mean μ_2 and variance σ_2^2.

We set the null hypothesis H_0 that there is no significant difference between the sample means *i.e.*, $\mu_1 = \mu_2$.

Under the null hypothesis H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

If the samples are drawn from the same population with common standard deviation (σ), then under the null hypothesis the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (\because \ \sigma_1 = \sigma_2 = \sigma)$$

Note. 1. If $\sigma_1 \neq \sigma_2$ and σ_1 and σ_2 are not known, the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

2. If common standard deviation (σ) is not known and $\sigma_1 = \sigma_2$ than σ can be obtained by using

$$\sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}}$$

The test statistic is $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

(*ii*) Standard Deviations. This test is used to test the significant difference between the standard deviations of two populations.

Let two independent random sample of sizes n_1 and n_2 having standard deviations s_1 and s_2 be drawn from the two normal population with standard deviation σ_1 and σ_2 respectively.

We set the null hypothesis H_0 that the sample standard deviations do not differ significantly *i.e.*, $\sigma_1 = \sigma_2$.

Under the null hypothesis H_0 the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

If σ_1 and σ_2 are unknown then the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}}.$$

Example 11.22. *Examine whether there is any significant difference between the two samples for the following data:*

Sample	*Size*	*Mean*
1	*50*	*140*
2	*60*	*150*

Standard deviation of the population = 10.

Solution. Here, $n_1 = 50$, $n_2 = 60$, $\bar{x}_1 = 140$, $\bar{x}_2 = 150$, $\sigma = 10$

Null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, samples are drawn from the same normal population.

NOTES

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

Under H_0 the test statistics is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{140 - 150}{10\sqrt{\frac{1}{50} + \frac{1}{60}}} = -\frac{10}{1.915} = -5.22$$

Since $|Z| = 5.22 > 3$, H_0 is rejected. Hence the samples are not drawn from the same normal population.

Example 23. *Intelligence tests on two groups of boys and girls gave the following results.*

	Mean	*S.D.*	*Size*
Girls	*70*	*10*	*70*
Boys	*75*	*11*	*100*

Examine if the difference between mean scores is significant.

Solution. Here, $n_1 = 70$, $n_2 = 100$, $\bar{x}_1 = 70$, $\bar{x}_2 = 75$, $s_1 = 10$, $s_2 = 11$

Null hypothesis H_0 : There is no significant difference between mean scores *i.e.*, $\bar{x}_1 = \bar{x}_2$.

Alternative hypothesis $H_1 : \bar{x}_1 \neq \bar{x}_2$ (two-tailed test)

Under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{70 - 75}{\sqrt{\frac{10^2}{70} + \frac{11^2}{100}}} = -\frac{5}{2.639} = -1.895$$

Since $|Z| = 1.895 < 1.96$, H_0 is accepted at 5% level of significance. Hence there is no significant difference between mean scores.

Example 11.24. *The means of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms respectively. Can the samples be regarded as drawn from the same population of standard deviation 6.25 cms?*

Solution. Here, $n_1 = 1000$, $n_2 = 2000$, $\bar{x}_1 = 168.75$, $\bar{x}_2 = 170$, $\sigma = 6.25$

Null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, samples are drawn from the same population.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (two-tailed test)

Under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{168.75 - 170}{6.25\sqrt{\frac{1}{1000} + \frac{1}{2000}}} = -\frac{1.25}{0.242} = -5.165$$

Since $|Z| = 5.165 > 1.96$, H_0 is rejected at 5% level of significance. Hence the samples are not drawn from the same population.

Example 11.25. *Two random samples of sizes 1000 and 2000 farms gave an average yield of 2000 kg and 2050 kg respectively. The variance of wheat farms in the country may be taken as 10 kg. Examine whether the two samples differ significantly in yield.*

Solution. Here, $n_1 = 1000$, $n_2 = 2000$, $\bar{x}_1 = 2000$, $\bar{x}_2 = 2050$, $\sigma^2 = 100$ *i.e.*, $\sigma = 10$

Null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, samples are drawn from the same population.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{2000 - 2050}{10\sqrt{\frac{1}{1000} + \frac{1}{2000}}} = -\frac{50}{0.387} = -129.20$$

Since $|Z| = 129.20 > 3$ (maximum value of Z), highly significant, H_0 is rejected. Hence the samples are not drawn from the same normal population.

Example 11.26. *Random samples drawn from two large cities gave the following information relating to the heights of adult males:*

	Mean height (in inches)	*Standard deviation*	*No. in samples*
City 1	*67.42*	*2.58*	*1000*
City 2	*67.25*	*2.50*	*1200*

Test the significance of difference in standard deviations of the samples at 5% level of significance.

Solution. Here, $n_1 = 1000$, $n_2 = 1200$, $\bar{x}_1 = 67.42$, $\bar{x}_2 = 67.25$, $s_1 = 2.58$, $s_2 = 2.50$, σ is not known.

Null hypothesis $H_0 : \sigma_1 = \sigma_2$ *i.e.*, the sample standard deviations do not differ significantly.

Alternative hypothesis $H_1 : \sigma_1 \neq \sigma_2$ (two-tailed test)

Under H_0 the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{2.58 - 2.50}{\sqrt{\frac{(2.58)^2}{2000} + \frac{(2.50)^2}{2400}}} = \frac{0.08}{0.077} = 1.039$$

Since $|Z| = 1.039 < 1.96$, H_0 is accepted. Hence sample standard deviations do not differ significantly.

Example 11.27. *In a survey of incomes of two classes of workers of two random samples gave the following data:*

	Size of sample	*Mean annual income in ₹*	*Standard deviation in ₹*
Sample 1	*100*	*582*	*24*
Sample 2	*100*	*546*	*28*

Examine whether the difference between

(i) Mean and

(ii) The standard deviations significant.

NOTES

Solution. Here, $n_1 = 100$, $n_2 = 100$, $\bar{x}_1 = 582$, $\bar{x}_2 = 546$, $s_1 = 24$, $s_2 = 28$

(*i*) Null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, sample means do not differ significantly.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{582 - 546}{\sqrt{\frac{(24)^2}{100} + \frac{(28)^2}{100}}} = \frac{36}{3.6878} = 9.762$$

Since $|Z| = 9.762 > 1.96$, highly significant, H_0 is rejected at 5% level of significance. Hence sample means differ significantly.

(*ii*) Null hypothesis $H_0 : \sigma_1 = \sigma_2$ *i.e.*, sample standard deviations do not differ significantly.

Alternative hypothesis $H_1 : \sigma_1 \neq \sigma_2$ (two-tailed test)

Under H_0 the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{24 - 28}{\sqrt{\frac{(24)^2}{200} + \frac{(28)^2}{200}}} = \frac{-4}{2.6077} = -1.53$$

Since $|Z| = 1.53 < 1.96$, H_0 is accepted at 5% level of significance. Hence sample standard deviations do not differ significantly.

EXERCISE 11.5

1. The number of accidents per day were studied for 144 days in city A and for 100 days in city B. The mean numbers of accidents and standard deviations were respectively 4.5 and 1.2 for city A and 5.4 and 1.5 for city B. Is city A more prone to accidents than city B.
2. The mean yields of a crop from two places in a district were 210 kgs and 220 kgs per acre from 100 acres and 150 acres respectively. Can it be regarded that the sample were drawn from the same district which has the standard deviation of 11 kgs per acre?
3. Given the following data:

	No. of cases	*Mean wages in* ₹	*Standard deviation of wages in* ₹
Sample 1	400	47.4	3.1
Sample 2	900	50.3	3.3

Examine whether the two mean wages differ significantly.

4. A sample of heights of 6400 soldiers has a mean of 67.85 inches and a standard deviation of 2.56 inches. While another sample of heights of 1600 sailors has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the sailors are on the average taller than soldiers?
5. Intelligence tests on two groups of boys and girls gave the following results:

	Mean	*S.D*	*Size*
Girls	75	8	60
Boys	73	10	100

Examine if the difference between mean scores is significant.

6. The yield of a crop in a random sample of 1000 farms in a certain area has a standard deviation of 192 kgs. Another random sample of 1000 farms gives a standard deviation of 224 kgs. Are the standard deviations significantly different?
7. The standard deviation of a random sample of 900 members is 4.6 and that of another random sample of 1600 is 4.8. Examine if the standard deviations are significantly different.
8. The mean yield of two sets of plots and their variability are as follow:

	Set of 40 plots	*Set of 60 plots*
Mean yield per plot	1258 kgs	1243 kgs
S.D. per plot	34	28

Examine whether

(*i*) the difference in the variability in yields is significant,

(*ii*) the difference in the mean yields is significant.

Answers

1. No
2. No
3. Yes, highly significant
4. Highly significant
5. Not significant at 5%
6. Yes
7. Not significant
8. (*i*) Not significant (*ii*) significant.

11.19. CHI-SQUARE TEST

In test of hypothesis of parameters, it is usually assumed that the random variable follows a particular distribution. To confirm whether our assumption is right, Chi-square test is used which measures the discrepancy between the observed (actual) frequencies and theoretical (expected) frequencies, on the basis of outcomes of a trial or observational data. Chi-square is a letter of the Greek alphabet and is denoted by χ^2. It is a continuous distribution which assumes only positive values.

11.20. CHI-SQUARE TEST TO TEST THE GOODNESS OF FIT

The value of χ^2 is used to test whether the deviations of the observed (actual) frequencies from the theoretical (expected) frequencies are significant or not. Chi-square test is also used to test whether a set of observations fit a given distribution or not. Therefore, chi-square provides a test of goodness of fit.

If $O_1, O_2, \ldots\ldots, O_n$ is a set of observed (actual) frequencies and $E_1, E_2, \ldots\ldots, E_n$ is the corresponding set of theoretical (expected) frequencies, then the statistic χ^2 is given by

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

is distributed with $(n - 1)$ degrees of freedom.

Here, we test the null hypothesis.

H_0 : There is no significant difference between the observed (actual) values and the corresponding expected (theoretical) values.

NOTES

v.s., H_1 : H_0 is not true.

If $\chi^2_{cal} \geq \chi^2_{tab}$ (or $\chi^2_{\alpha, n-1}$) then H_0 is rejected otherwise H_0 is accepted.

Note. If the null hypothesis H_0 is true, the test statistic χ^2 follow chi-square distribution with $(n-1)$ degrees of freedom, where

$$\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i \; ; \quad i.e. \quad \sum_{i=1}^{n} (O_i - E_i) = 0.$$

11.21. CHI-SQUARE TEST TO TEST THE INDEPENDENCE OF ATTRIBUTES

The value of χ^2 is used to test whether two attributes are associated or not, *i.e.* independence of attributes. To test the independence of attributes contingency table is used.

A contingency table is a two-way table in which rows are classified according to one attribute or criterion and columns are classified according to the other attribute or criterion. Each cell contains that number of items O_{ij} possessing the qualities of the ith row and jth column, where $i = 1, 2, \ldots\ldots, r$ and $j = 1, 2, \ldots\ldots, s$. In such a case contingency table is said to be of order $(r \times s)$. Each row or column total is known as marginal total. Also we have the sum of row totals $\sum_{i=1}^{r} R_i$ is equal to the sum of column totals $\sum_{j=1}^{s} C_j$, *i.e.*

$$\sum_{i} R_i = \sum_{j} C_j = N, \text{ where } N \text{ is the total frequency.}$$

Let us consider the two attributes A and B, where A divided into r classes $A_1, A_2, \ldots\ldots, A_r$ and B divided into s classes $B_1, B_2, \ldots\ldots, B_s$. If R_i represents the number of persons possessing the attributes A_i; C_j represents the number of persons possessing the attributes B_j and O_{ij} represent the number of persons possessing attributes A_i and B_j respectively. The contingency table of order $(r \times s)$ is shown in the following table:

Columns / Rows	B_1	B_2	.....	B_s	Total
A_1	O_{11}	O_{12}	.....	O_{1s}	R_1
A_2	O_{21}	O_{22}	.....	O_{2s}	R_2
⋮	⋮	⋮		⋮	⋮
A_r	O_{r1}	O_{r2}	.....	O_{rs}	R_r
Total	C_1	C_2	.....	C_s	N

Corresponding to each O_{ij} the expected frequency E_{ij} in a contingency table is calculated by

$$E_{ij} = \frac{R_i \times C_j}{N} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Here, we test the null hypothesis.

H_0 : There is no association between the attributes under study, *i.e.* attributes A and B are independent.

v.s., H_1 : attributes are associated, *i.e.*, attributes A and B are not independent.

H_0 can be tested by the statistic

$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ is distributed with $(r-1)(s-1)$ degrees of freedom.

If $\chi_{cal}^2 \geq \chi_{tab}^2$ (or $\chi^2_{\alpha,(r-1)(s-1)}$), then H_0 is rejected otherwise H_0 is accepted.

Note 1. For a contingency table with r rows and s columns, the degrees of freedom $= (r-1)(s-1)$.

2. For a 2×2 contingency table $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$ we use the following formula to calculate the value of statistic χ^2 as

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(b+d)(a+c)(c+d)},$$

where $N = a + b + c + d$

χ^2 has $(2-1)(2-1) = 1$ degree of freedom.

3. Yate's correction. In a 2×2 contingency table, if any of cell frequency is less than 5, we make a correction to make χ^2 continuous. Decrease by $\frac{1}{2}$ those cell frequencies which are greater than expected frequencies and increase by $\frac{1}{2}$ those cell frequencies which are less than expected frequencies. This will affect the marginal totals. This correction is known a Yate's correction.

After applying the Yate's correction, the corrected value of χ^2 is given by

$$\chi^2 = \frac{N\left(|ad - bc| - \frac{N}{2}\right)^2}{(a+b)(b+d)(a+c)(c+d)}$$

11.22. CONDITIONS FOR χ^2 TEST

1. The number of observations collected must be large, *i.e.* $n \geq 30$.
2. No theoretical frequency should be very small.
3. The sample observations should be independent.
4. N, the total of frequencies should be reasonably large, say, greater than 50.

NOTES

11.23. USES OF χ^2 TEST

1. To test the goodness of fit.
2. To test the discrepancies between observed and expected frequencies.
3. To determine the association between attributes.

Example 11.28. *The following table gives the number of accidents that took place in an industry during various days of the week. Test whether the accidents are uniformly distributed over the week.*

Days	*Mon.*	*Tue.*	*Wed.*	*Thu.*	*Fri.*	*Sat.*
No. of accidents	*16*	*20*	*14*	*13*	*17*	*16*

Solution. Here, $n = 6$, total number of accidents = 96

Null hypothesis H_0 : the accidents are uniformly distributed over the week.

Under H_0, the expected number of accidents of each of these days

$$= \frac{\text{Total no. of accidents}}{\text{No. of days}} = \frac{96}{6} = 16$$

The observed and expected number of accidents are given below:

O_i	16	20	14	13	17	16
E_i	16	16	16	16	16	16
$(O_i - E_i)^2$	0	16	4	9	1	0

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{0 + 16 + 4 + 9 + 1 + 0}{16} = \frac{30}{16} = 1.875.$$

Tabulated value of χ^2 for 5 (6 – 1 = 5) degrees of freedom at 5% level of significance is 11.07.

Since calculated value of χ^2 is less than tabulated value of χ^2, so H_0 is accepted, *i.e.,* the accidents are uniformly distributed over the week.

Example 11.29. *A die is thrown 120 times and the result of these throws are given as:*

No. appeared on the die	*1*	*2*	*3*	*4*	*5*	*6*
Frequency	*16*	*30*	*22*	*18*	*14*	*20*

Test whether the die is biased or not.

Solution. Here, $n = 6$, total frequency = 120

Null hypothesis H_0 : die is unbiased

Under H_0, the expected frequencies for each digit = $\frac{120}{6} = 20$

The observed and expected frequencies are given below :

O_i	16	30	22	18	14	20
E_i	20	20	20	20	20	20
$(O_i - E_i)^2$	16	100	4	4	36	0

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{16 + 100 + 4 + 4 + 36 + 0}{20} = \frac{160}{20} = 8$$

Tabulated value of χ^2 for 5 (6 – 1 = 5) degrees of freedom at 5% level of significance is 11.07. Since calculated value of χ^2 is less than tabulated value of χ^2, so H_0 is accepted, *i.e.* the die is unbiased.

Example 11.30. *The following table shows the distribution of digits in numbers chosen at random from a telephone directory.*

Digits	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Frequency	*1026*	*1107*	*997*	*966*	*1075*	*933*	*1107*	*972*	*964*	*853*

Test at 5% level whether the digits may be taken to occur equally frequently in the directory.

Solution. Here, $n = 10$, total frequency = 10,000

Null hypothesis H_0 : all the digits occur equally frequently in the directory

Under H_0, the expected frequency of each of the digits $= \frac{10,000}{10} = 1000$

The observed and expected frequencies are given below:

O_i	1026	1107	997	966	1075	933	1107	972	964	853
E_i	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
$(O_i - E_i)^2$	676	11449	9	1156	5625	4489	11449	784	1296	21609

$$\chi^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} = \frac{676 + 11449 + \ldots\ldots + 21609}{1000}$$

$$= \frac{58542}{1000} = 58.542$$

Tabulated value of χ^2 for 9 (10 – 1 = 9) degrees of freedom at 5% level of significance is 16.92.

Since calculated value of χ^2 is greater than tabulated value of χ^2, so H_0 is rejected, *i.e.*, all the digits in the numbers in the telephone directory do not occur equally frequently.

Example 11.31. *Fit a Poisson distribution for the following data and test the goodness of fit.*

No. of defects (x)	*0*	*1*	*2*	*3*	*4*	*5*
Frequency	*6*	*13*	*13*	*8*	*4*	*3*

NOTES

Solution. Null hypothesis H_0 : Poisson distribution is a good fit to the data.

We first find the Poisson distribution for the above data.

Mean of given distribution $= \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{94}{47} = 2$

Here, $\lambda = 2$ (For a Poisson distribution mean $= \lambda$)

$N = \Sigma f_i = 47$

The expected frequencies of the Poisson distribution are given by

$$E(r) = N \times e^{-\lambda} \frac{\lambda^r}{r!} = 47 \times e^{-2} \frac{2^r}{r!}; r = 0, 1, 2, 3, 4, 5$$

The expected frequencies are as:

$$E(0) = 47 \times e^{-2} \frac{2^0}{0!} = 6.36 \approx 6 \quad (e^{-2} = 0.1353)$$

$$E(1) = 47 \times e^{-2} \frac{2^1}{1!} = 12.72 \approx 13$$

$$E(2) = 47 \times e^{-2} \frac{2^2}{2!} = 12.72 \approx 13$$

$$E(3) = 47 \times e^{-2} \frac{2^3}{3!} = 8.48 \approx 9$$

$$E(4) = 47 \times e^{-2} \frac{2^4}{4!} = 4.24 \approx 4$$

$$E(5) = 47 \times e^{-2} \frac{2^5}{5!} = 1.696 \approx 2$$

x	0	1	2	3	4	5
O_i	6	13	13	8	4	3
E_i	6.36	12.72	12.72	8.48	4.24	1.696
$(O_i - E_i)^2$	0.1296	0.0784	0.0784	0.2304	0.0576	1.7004

$$\chi^2 = \sum_{i=0}^{5} \frac{(O_i - E_i)^2}{E_i} = \frac{0.1296}{6.36} + \frac{0.0784}{12.72} + \frac{0.0784}{12.72} + \frac{0.2304}{8.48} + \frac{0.0576}{4.24} + \frac{1.7004}{1.696}$$

$$= 0.02038 + 0.00616 + 0.00616 + 0.02717 + 0.01358 + 1.0026$$

$$= 1.07605$$

Tabulated value of χ^2 for 4 $(6 - 2 = 4)$ degrees of freedom at 5% level of significance is 9.488.

Since calculated value of χ^2 is less than tabulated value of χ^2, so H_0 is accepted, *i.e.*, Poisson distribution is a good fit to the data.

Example 11.32. *Find the expected frequencies of 2 × 2 contingency table* $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$.

Solution.

Attributes	B_1	B_2	*Total*
A_1	a	b	$a + b$
A_2	c	d	$c + d$
Total	$a + c$	$b + d$	$N = a + b + c + d$

The expected frequencies are

$$E(a) = E(A_1, B_1) = \frac{(a+b)(a+c)}{a+b+c+d}$$

$$E(b) = E(A_1, B_2) = \frac{(a+b)(b+d)}{a+b+c+d}$$

$$E(c) = E(A_2, B_1) = \frac{(c+d)(a+c)}{a+b+c+d}$$

$$E(d) = E(A_2, B_2) = \frac{(c+d)(b+d)}{a+b+c+d}$$

Example 11.33. *In a locality 100 persons were randomly selected and asked about their educational achievements. The results are given below:*

Sex	*Education*		
	Middle	*High school*	*College*
Male	*10*	*15*	*25*
Female	*25*	*10*	*15*

Based on this information can you say the education depends on sex.

Solution. Null hypothesis H_0: Education is independent of sex.

Under the null hypothesis expected frequencies can be calculated by using

$$E_{ij} = \frac{R_i \times C_j}{N}$$

$$(i = 1, 2; j = 1, 2, 3)$$

Sex	*Education*			*Total*
	Middle	*High school*	*College*	
Male	10	15	25	50 (R_1)
Female	25	10	15	50 (R_2)
Total	35 (C_1)	25 (C_2)	40 (C_3)	N = 100

Expected frequencies are:

Sex	*Education*			*Total*
	Middle	*High school*	*College*	
Male	$\frac{50 \times 35}{100} = 17.5$	$\frac{50 \times 25}{100} = 12.5$	$\frac{50 \times 40}{100} = 20$	50
Female	$\frac{50 \times 35}{100} = 17.5$	$\frac{50 \times 25}{100} = 12.5$	$\frac{50 \times 40}{100} = 20$	50
Total	35	25	40	100

NOTES

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \frac{(10-17.5)^2}{17.5} + \frac{(15-12.5)^2}{12.5} + \frac{(25-20)^2}{20} + \frac{(25-17.5)^2}{17.5} + \frac{(10-12.5)^2}{12.5} + \frac{(15-20)^2}{20}$$

$$= 3.214 + 0.5 + 1.25 + 3.214 + 0.5 + 1.25 = 9.928$$

Tabulated value of χ^2 for 2 [(2 – 1) (3 – 1) =2] degrees of freedom at 5% level of significance is 5.991. Since calculated value of χ^2 is greater than tabulated value of χ^2, so H_0 is rejected, *i.e.*, education is not independent of sex or there is a relation between education and sex.

Example 11.34. *The following table gives the number of good and bad parts produced by each of the three shifts in a factory.*

	Good parts	*Bad parts*	*Total*
Day shift	*960*	*40*	*1000*
Evening shift	*940*	*50*	*990*
Night shift	*950*	*45*	*995*
Total	*2850*	*135*	*2985*

Test whether the production of bad parts is independent of the shifts on which they were produced.

Solution. Null hypothesis H_0 : The production of bad parts is independent of the shift on which they were produced, *i.e.* production and shifts are independent.

Under the null hypothesis expected frequencies can be calculated by using

$$E_{ij} = \frac{R_i \times C_j}{N} \quad (i = 1, 2, 3; j = 1, 2)$$

Expected frequencies are:

	Good parts	*Bad parts*	*Total*
Day shift	$\frac{1000 \times 2850}{2985} = 954.774$	$\frac{1000 \times 135}{2985} = 45.226$	1000
Evening shift	$\frac{990 \times 2850}{2985} = 945.226$	$\frac{990 \times 135}{2985} = 44.774$	990
Night shift	$\frac{995 \times 2850}{2985} = 950.000$	$\frac{995 \times 135}{2985} = 45.000$	995
	2850	135	2985

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \frac{(960-954.774)^2}{954.774} + \frac{(40-45.226)^2}{45.226} + \frac{(940-945.226)^2}{945.226} + \frac{(50-44.774)^2}{44.774} + \frac{(950-950)^2}{950} + \frac{(45-45)^2}{45}$$

$$= 0.0286 + 0.6039 + 0.0289 + 0.6099 + 0 + 0 = 1.2713$$

Tabulated value of χ^2 for 2 [(3 – 1) (2 – 1) = 2] degrees of freedom at 5% level of significance is 5.991

Since calculated value of χ^2 is less than tabulated value of χ^2, so H_0 is accepted, *i.e.*, the production of bad parts is independent of the shift on which they were produced.

11.24. SUMMARY

- A statistical measure based only on all the units selected in a sample is called 'statistic', *e.g.*, sample mean, sample standard deviation, proportion of defectives, etc. whereas a statistical measure based on all the units in the population is called 'parameter'. The terms like mean, median, mode, standard deviation are called parameters when they describe the characteristics of the population and are called statistic when they describe the characteristics of the sample.
- A statistical hypothesis is a statement about a population parameter. There are two types of statistical hypothesis, null hypothesis and alternative hypothesis.
- The hypothesis formulated for the sake of rejecting it under the assumption that it is true, is called the null hypothesis and is denoted by H_0. Null hypothesis asserts that there is no significant difference between the sample statistic and the population parameter and whatever difference is observed that is merely due to fluctuations in sampling from the same population.
- The number of independent variates which make up the statistic is known as the degree of freedom (d.f.) and is denoted by ν (the letter 'Nu' of the Greek alphabet).
- In test of hypothesis of parameters, it is usually assumed that the random variable follows a particular distribution. To confirm whether our assumption is right, Chi-square test is used which measures the discrepancy between the observed (actual) frequencies and theoretical (expected) frequencies, on the basis of outcomes of a trial or observational data. Chi-square is a letter of the Greek alphabet and is denoted by χ^2. It is a continuous distribution which assumes only positive values.

11.25. REVIEW EXERCISES

1. The frequency distribution of the digits on a set of random numbers was observed to be:

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Test the hypothesis that the digits are uniformly distributed.

2. The following table gives the number of accidents that took place in an industry during various days of the week:

Days	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
No. of accidents	14	18	12	11	15	14

Test if accidents are uniformly distributed over the week.

NOTES

3. A die is thrown 276 times and the results of these throws are given below:

No. appeared on the die	1	2	3	4	5	6
Frequency	40	32	29	59	57	59

Test whether the die is biased or not.

4. A sample analysis of examination results of 500 students was made. It was found that 220 had failed; 170 had secured a third class; 90 were placed in second class; 20 got first class. Are these results commensurable with the general examination result which is in the ratio of 4 : 3 : 2 : 1 for the above said categories respectively.

5. Four dice were thrown 112 times and the number of times 1, 3 or 5 was thrown were as under:

No. of dice throwing 1, 3 or 5	0	1	2	3	4
Frequency	10	25	40	30	7

Test the hypothesis that all dice were fair.

6. Fit a Poisson distribution for the following data and test the goodness of fit.

No. of defects (x)	0	1	2	3	4
Frequency	109	65	22	3	1

7. For the data given in the following table use χ^2-test to test the effectiveness of inoculation in preventing the attack of smallpox.

	Attacked	*Not attacked*
Inoculated	25	220
Not inoculated	90	160

8. Two investigators draw samples from the same town in order to estimate the number of persons falling in the income groups 'poor', middle class' and 'well to do'. Their results are as follows:

Investigator	*Income groups*		
	Poor	*Middle class*	*Well to do*
A	140	100	15
B	140	50	20

Test whether the sampling techniques of the two investigators are significantly dependent of the income groups of people.

Answers

1. Yes
2. Yes
3. Biased
4. No
5. Yes
6. Poisson distribution is a good fit to the data.
7. Inoculation against smallpox is a preventive measure.
8. Sampling techniques are dependent of the income groups.