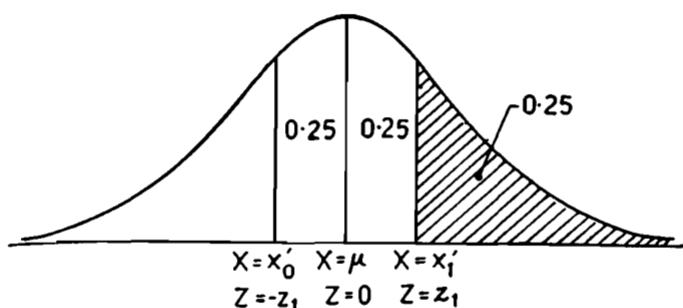


From (\*), obviously the points  $x_0'$  and  $x_1'$  are located as shown in the figure.



$$\text{When } X = x_1', \quad Z = \frac{x_1' - 12}{4} = z_1 \text{ (say)}$$

$$\text{and when } X = x_0', \quad Z = \frac{x_0' - 12}{4} = -z_1 \quad (\text{It is obvious from the figure})$$

We have

$$P(Z > z_1) = 0.25 \Rightarrow P(0 < Z < z_1) = 0.25$$

$$\therefore z_1 = 0.67 \quad (\text{From tables})$$

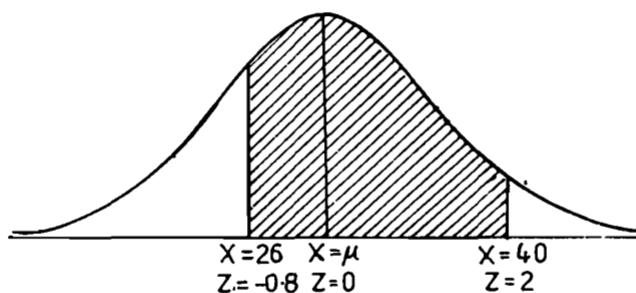
$$\text{Hence } \frac{x_1' - 12}{4} = 0.67 \Rightarrow x_1' = 12 + 4 \times 0.67 = 14.68$$

$$\text{and } \frac{x_0' - 12}{4} = -0.67 \Rightarrow x_0' = 12 - 4 \times 0.67 = 9.32$$

**Example 8.13.**  $X$  is a normal variate with mean 30 and S.D. 5. Find the probabilities that

- (i)  $26 \leq X \leq 40$ , (ii)  $X \geq 45$ , and (iii)  $|X - 30| > 5$ .

**Solution.** Here  $\mu = 30$  and  $\sigma = 5$ .



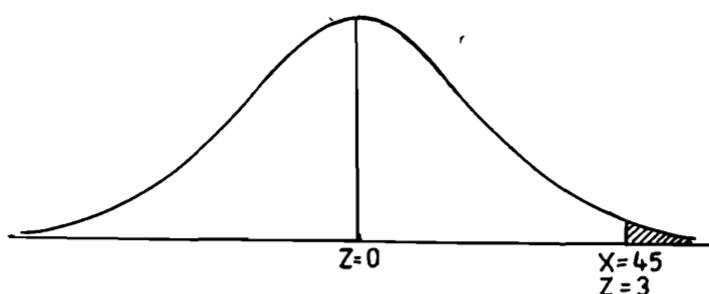
$$(i) \text{ When } X = 26, Z = \frac{X - \mu}{\sigma} = \frac{26 - 30}{5} = -0.8$$

and when

$$X = 40, Z = \frac{40 - 30}{5} = 2$$

$$\begin{aligned}\therefore P(26 \leq X \leq 40) &= P(-0.8 \leq Z \leq 2) \\&= P(-0.8 \leq Z \leq 0) + P(0 \leq Z \leq 2) \\&= P(-0.8 \leq Z \leq 0) + 0.4772 \quad (\text{From tables}) \\&= P(0 \leq Z \leq 0.8) + 0.4772 \quad (\text{From symmetry}) \\&= 0.2881 + 0.4772 = 0.7653\end{aligned}$$

$$P(X \geq 45) = ?$$



$$\text{When } X = 45, \quad Z = \frac{45 - 30}{5} = 3$$

$$\therefore P(X \geq 45) = P(Z \geq 3) = 0.5 - P(0 \leq Z \leq 3) \\= 0.5 - 0.49865 = 0.00135$$

$$\begin{aligned}(iii) \quad P(|X - 30| \leq 5) &= P(25 \leq X \leq 35) = P(-1 \leq Z \leq 1) \\&= 2P(0 \leq Z \leq 1) = 2 \times 0.3413 = 0.6826 \\ \therefore P(|X - 30| > 5) &= 1 - P(|X - 30| \leq 5) \\&= 1 - 0.6826 = 0.3174\end{aligned}$$

**Example 8.14.** The mean yield for one-acre plot is 662 kilos with a s.d. 32 kilos. Assuming normal distribution, how many one-acre plots in a batch of 1,000 plots would you expect to have yield (i) over 700 kilos, (ii) below 650 kilos, and (iii) what is the lowest yield of the best 100 plots?

**Solution.** If the r.v.  $X$  denotes the yield (in kilos) for one-acre plot, then we are given that  $X \sim N(\mu, \sigma^2)$ , where  $\mu = 662$  and  $\sigma = 32$ .

(i) The probability that a plot has a yield over 700 kilos is given by

$$\begin{aligned}P(X > 700) &= P(Z > 1.19); \quad Z = \frac{X - 662}{32} \\&= 0.5 - P(0 \leq Z \leq 1.19) \\&= 0.5 - 0.3830 \\&= 0.1170\end{aligned}$$

Hence in a batch of 1,000 plots, the expected number of plots with yield over 700 kilos is  $1,000 \times 0.117 = 117$ .

(ii) Required number of plots with yield below 650 kilos is given by

$$\begin{aligned}
 1000 \times P(X < 650) &= 1000 \times P(Z < -0.38) \\
 &= 1000 \times P(Z > 0.38) \\
 &= 1000 \times [0.5 - P(0 \leq Z \leq 0.38)] \\
 &= 1000 \times [0.5 - 0.1480] = 1000 \times 0.352 \\
 &= 352
 \end{aligned}
 \quad \left[ \begin{array}{l} Z = \frac{650 - 662}{32} \\ \text{(By symmetry)} \end{array} \right]$$

(iii) The lowest yield, say,  $x_1$  of the best 100 plots is given by

$$P(X > x_1) = \frac{100}{1000} = 0.1$$

$$\text{When } X = x_1, Z = \frac{x_1 - \mu}{\sigma} = \frac{x_1 - 662}{32} = z_1 \text{ (say)} \quad \dots (*)$$

$$\text{such that } P(Z > z_1) = 0.1 \Rightarrow P(0 \leq Z \leq z_1) = 0.4.$$

$$\Rightarrow z_1 = 1.28 \text{ (approx.)} \quad [\text{From Normal Probability Tables}]$$

Substituting in (\*), we get

$$\begin{aligned}
 x_1 &= 662 + 32, z = 662 + 32 \times 1.28 \\
 &= 662 + 40.96 = 702.96
 \end{aligned}$$

Hence the best 100 plots have yield over 702.96 kilos.

**Example 8-15.** There are six hundred Economics students in the post-graduate classes of a university, and the probability for any student to need a copy of a particular book from the university library on any day is 0.05. How many copies of the book should be kept in the university library so that the probability may be greater than 0.90 that none of the students needing a copy from the library has to come back disappointed? (Use normal approximation to the binomial distribution). [Delhi Univ. M.A. (Eco.), 1989]

**Solution.** We are given :

$$n = 600, p = 0.05, \mu = np = 600 \times 0.05 = 30$$

$$\sigma^2 = npq = 600 \times 0.05 \times 0.95 = 28.5 \Rightarrow \sigma = \sqrt{28.5} = 5.3$$

We want  $x_1$  such that

$$P(X < x_1) > 0.90$$

$$\Rightarrow P(Z < z_1) > 0.90$$

$$\left[ z_1 = \frac{x_1 - 30}{5.3} \right]$$

$$\Rightarrow P(0 < Z < z_1) > 0.40$$

[From Normal Probability Tables]

$$\Rightarrow z_1 > 1.28$$

$$\Rightarrow \frac{x_1 - 30}{5.3} > 1.28 \Rightarrow x_1 > 30 + 5.3 \times 1.28$$

$$\Rightarrow x_1 > 30 + 6.784 \Rightarrow x_1 > 36.784 \approx 37$$

Hence the university library should keep at least 37 copies of the book.

**Example 8-16.** The marks obtained by a number of students for a certain subject are assumed to be approximately normally distributed with mean value 65

and with a standard deviation of 5. If 3 students are taken at random from this set what is the probability that exactly 2 of them will have marks over 70?

**Solution.** Let the r.v.  $X$  denote the marks obtained by the given set of students in the given subject. Then we are given that  $X \sim N(\mu, \sigma^2)$  where  $\mu = 65$  and  $\sigma = 5$ . The probability ' $p$ ' that a randomly selected student from the given set gets marks over 70 is given by

$$p = P(X > 70)$$

$$\text{When } X = 70, Z = \frac{X - \mu}{\sigma} = \frac{70 - 65}{5} = 1.$$

$$\begin{aligned} \therefore p &= P(X > 70) = P(Z > 1) \\ &= 0.5 - P(0 \leq Z \leq 1) \\ &= 0.5 - 0.3413 = 0.1587 \quad [\text{From Normal probability tables}] \end{aligned}$$

Since this probability is same for each student of the set, the required probability that 'out of 3 students selected at random from the set, exactly 2 will have marks over 70', is given by the binomial probability law:

$${}^3C_2 p^2 \cdot (1-p)^1 = 3 \times (0.1587)^2 \times (0.8413) = 0.06357$$

**Example 8-17.** (a) If  $\log_{10} X$  is normally distributed with mean 4 and variance 4, find the probability of

$$1.202 < X < 83180000$$

(Given  $\log_{10} 1202 = 3.08$ ,  $\log_{10} 8318 = 3.92$ ).

(b)  $\log_{10} X$  is normally distributed with mean 7 and variance 3,  $\log_{10} Y$  is normally distributed with mean 3 and variance unity. If the distributions of  $X$  and  $Y$  are independent, find the probability of  $1.202 < (X/Y) < 83180000$ .

[Given  $\log_{10} (1202) = 3.08$ ,  $\log_{10} (8318) = 3.92$ ]

**Solution.** (a) Since  $\log X$  is a non-decreasing function of  $X$ , we have

$$\begin{aligned} P(1.202 < X < 83180000) &= P(\log_{10} 1.202 < \log_{10} X < \log_{10} 83180000) \\ &= P(0.08 < \log_{10} X < 7.92) \\ &= P(0.08 < Y < 7.92) \end{aligned}$$

where  $Y = \log_{10} X \sim N(4, 4)$  (given).

$$\text{When } Y = 0.08, Z = \frac{0.08 - 4}{2} = -1.96$$

$$\text{and when } Y = 7.92, Z = \frac{7.92 - 4}{2} = 1.96$$

$$\begin{aligned} \therefore \text{Required probability} &= P(0.08 < Y < 7.92) \\ &= P(-1.96 < Z < 1.96) = 2P(0 < Z < 1.96) \\ &\quad (\text{By symmetry}) \\ &= 2 \times 0.4750 = 0.9500 \end{aligned}$$

$$(b) P[1.202 < (X/Y) < 83180000]$$

$$\begin{aligned} &= P[\log_{10} 1.202 < \log_{10}(X/Y) < \log_{10} 83180000] \\ &= (0.08 < U < 7.92) \end{aligned}$$

$$\text{where } U = \log_{10}(X/Y) = \log_{10} X - \log_{10} Y$$

Since  $\log_{10} X \sim N(7, 3)$  and  $\log_{10} Y \sim N(3, 1)$ , are independent,

$$\log_{10} X - \log_{10} Y \sim N(7-3, 3+1) \quad (\text{c.f. Remark 1, § 8.2.8})$$

$$\Rightarrow U = (\log_{10} X - \log_{10} Y) \sim N(4, 4)$$

∴ Required probability is given by

$$P(U < 0.08) = P(0.08 < U < 7.92), \text{ where } U \sim N(4, 4)$$

$$= 0.95$$

[See part (a)]

**Example 8.18.** Two independent random variates  $X$  and  $Y$  are both normally distributed with means 1 and 2 and standard deviations 3 and 4 respectively. If  $Z = X - Y$ , write the probability density function of  $Z$ . Also state the median, s.d. and mean of the distribution of  $Z$ . Find  $\text{Prob. } \{Z + 1 \leq 0\}$ .

**Solution.** Since  $X \sim N(1, 9)$  and  $Y \sim N(2, 16)$  are independent,  $Z = X - Y \sim N(1-2, 9+16)$ , i.e.,  $Z = X - Y \sim N(-1, 25)$ . Hence p.d.f. of  $Z$  is

$$p(z) = \frac{1}{5\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z+1}{5}\right)^2\right]; -\infty < z < \infty.$$

For the distribution of  $Z$ ,

$$\text{Median} = \text{Mean} = -1 \quad \text{and} \quad \text{s.d.} = \sqrt{25} = 5$$

$$P(Z + 1 \leq 0) = P(Z \leq -1)$$

$$= P(U \leq 0); \quad \left[ U = \frac{Z+1}{5} \sim N(0, 1) \right]$$

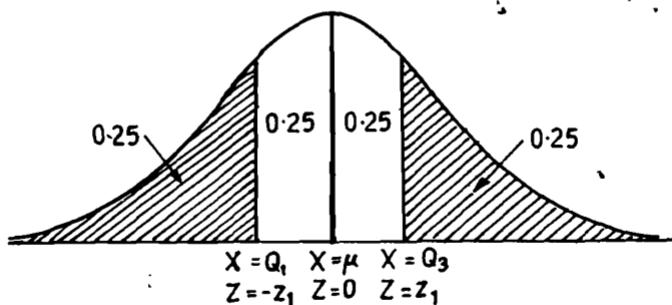
$$= 0.5$$

**Example 8.19.** Prove that for the normal distribution, the quartile deviation, the mean deviation and standard deviation are approximately 10 : 12 : 15. [Dibrugarh Univ. B.Sc. 1993]

**Solution.** Let  $X$  be a  $N(\mu, \sigma^2)$ . If  $Q_1$  and  $Q_3$  are the first and third quartiles respectively, then by definition

$$P(X < Q_1) = 0.25 \quad \text{and} \quad P(X > Q_3) = 0.25$$

The points  $Q_1$  and  $Q_3$  are located as shown in the figure given below.



When  $X = Q_3, Z = \frac{Q_3 - \mu}{\sigma} = z_1$ , (say),

and when  $X = Q_1, Z = \frac{Q_1 - \mu}{\sigma} = -z_1$  (This is obvious from the figure)

Subtracting, we have

$$\frac{Q_3 - Q_1}{\sigma} = 2z_1$$

The quartile deviation is given by

$$Q.D. = \frac{Q_3 - Q_1}{2} = \sigma z_1$$

From the figure, obviously, we have

$$P(0 < Z < z_1) = 0.25 \Rightarrow z_1 = 0.67 \text{ (approx.)} \quad (\text{From Normal Tables})$$

$$\therefore Q.D. = \sigma z_1 = 0.67 \sigma = \frac{2}{3} \sigma$$

For normal distribution mean deviation about mean (c.f. § 8-2-10) is given by

$$M.D. = \sqrt{2/\pi} \sigma = \frac{4}{5} \sigma$$

$$\text{Hence Q.D. : M.D. : S.D. :: } \frac{2}{3} \sigma : \frac{4}{5} \sigma : \sigma :: \frac{2}{3} : \frac{4}{5} : 1 :: 10 : 12 : 15$$

**Example 8-20 (a).** In a distribution exactly normal, 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution? [Kerala Univ. B.Sc., May 1991]

(b) Of a large group of men, 5% are under 60 inches in height and 40% are between 60 and 65 inches. Assuming a normal distribution, find the mean height and standard deviation. [Nagpur Univ. B.Sc., 1992]

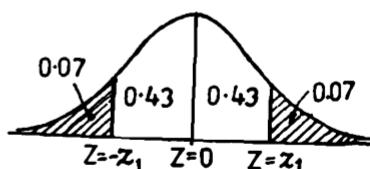
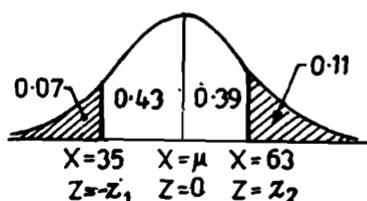
**Solution.** If  $X \sim N(\mu, \sigma^2)$ , then we are given

$$P(X < 35) = 0.07 \Rightarrow P(X > 63) = 0.11 \text{ and } P(X < 35) = 0.07$$

The points  $X = 63$  and  $X = 35$  are located as shown in Fig. (i) below.

Since the value  $X = 35$  is located to the left of the ordinate at  $X = \mu$ , the corresponding value of  $Z$  is negative.

When  $X = 35, Z = \frac{35 - \mu}{\sigma} = -z_1$ , (say),



and when  $X = 63$ ,  $Z = \frac{63 - \mu}{\sigma} = z_2$ , (say).

Thus we have, as is obvious from figures (i) and (ii)

$$P(0 < Z < z_2) = 0.39 \text{ and } P(0 < Z < z_1) = 0.43$$

Hence from normal tables, we have

$$z_2 = 1.23 \text{ and } z_1 = 1.48$$

$$\therefore \frac{63 - \mu}{\sigma} = 1.23 \text{ and } \frac{35 - \mu}{\sigma} = -1.48$$

Subtracting, we get

$$\frac{28}{\sigma} = 2.71 \Rightarrow \sigma = \frac{28}{2.71} = 10.33$$

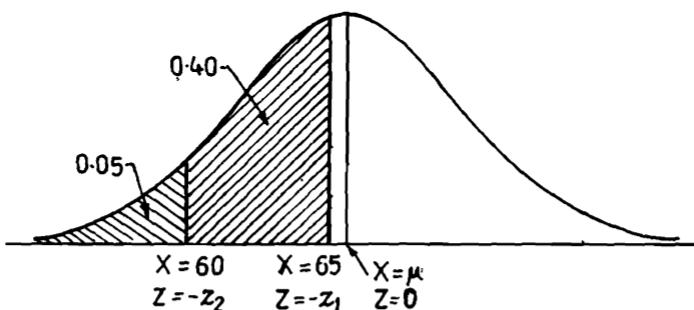
$$\therefore \mu = 35 + 1.48 \times 10.33 = 35 + 15.3 = 50.3$$

(b) We are given

$$P(X < 60) = 0.05 \text{ and } P(60 < X < 65) = 0.40$$

$$\text{i.e., } P(X < 65) = 0.45$$

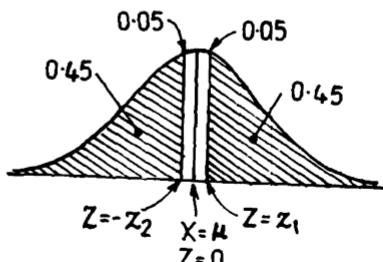
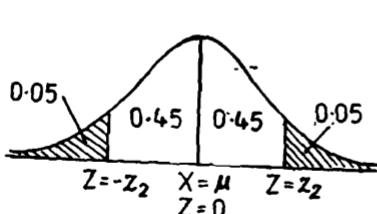
Since the total area to the left of the ordinate at  $X = \mu$  is 0.5, both the points  $X = 60$  and  $X = 65$  are located to the left of  $X = \mu$  and consequently the corresponding values of  $Z$  are negative.



Let  $X \sim N(\mu, \sigma^2)$ .

$$\text{When } X = 65, \quad Z = \frac{65 - \mu}{\sigma} = -z_1 \text{ (say),}$$

$$\text{and when } X = 60, \quad Z = \frac{60 - \mu}{\sigma} = -z_2 \text{ (say).}$$



Thus we have

$$P(0 < Z < z_2) = 0.45 \text{ and } P(0 < Z < z_1) = 0.05$$

$\therefore z_2 = 1.645$  and  $z_1 = 0.13$  (approx.) (From Normal Tables)

$$\text{Hence } \frac{60 - \mu}{\sigma} = -1.645 \dots (*) ; \text{ and } \frac{65 - \mu}{\sigma} = -0.13 \dots (**)$$

$$\text{Dividing, we get } \frac{60 - \mu}{65 - \mu} = \frac{1.645}{0.13} \Rightarrow \mu = \frac{19825}{303} = 65.42$$

$$\therefore \text{From } (*), \text{ we have } \sigma = \frac{60 - 65.42}{-1.645} = 3.29$$

**Remarks.** If we substitute the value of  $\mu$  in (\*\*), we get  $\sigma = 3.23$  which is only an approximate value since the value of  $z_1 = 0.13$ , seen from the table, is not exact but only approximate. On the other hand, the value of  $z_2 = 1.645$  is exact and hence use of (\*) for estimating  $\sigma$  gives better approximation.

**Example 8.21** If the skulls are classified as A, B and C according as the length-breadth index is under 75, between 75 and 80, or over 80, find approximately (assuming that the distribution is normal) the mean and standard deviation of a series in which A are 58%, B are 38% and C are 4%, being given that if

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_0^t \exp(-x^2/2) dx,$$

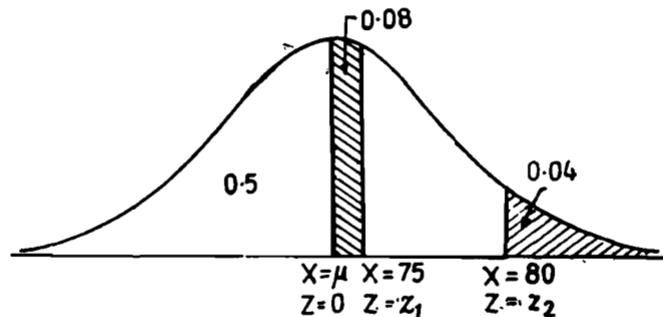
$$\text{then } f(0.20) = 0.08 \text{ and } f(1.75) = 0.46$$

[Delhi Univ. B.Sc., 1989; Burdwan Univ. B.Sc., 1990]

**Solution.** Let the length-breadth index be denoted by the variable  $X$ , then we are given

$$P(X < 75) = 0.58 \text{ and } P(X > 80) = 0.04 \quad \dots (1)$$

Since  $P(X < 75)$  represents the total area to the left of the ordinate at the point  $X = 75$  and  $P(X > 80)$  represents the total area to the right of the ordinate at the point  $X = 80$ , it is obvious from (1) that the points  $X = 75$  and  $X = 80$  are located at the positions shown in the figure below.



Now  $\frac{1}{\sqrt{2}\pi} \int_0^t \exp(-x^2/2) dx$  represents the area under standard normal curve between the ordinates at  $Z = 0$  and  $Z = t$ ,  $Z$  being a  $N(0, 1)$  variate.

$$\therefore f(t) = \frac{1}{\sqrt{2}\pi} \int_0^t \exp(-x^2/2) dx = P(0 < Z < t)$$

Hence  $f(0.20) = P(0 < Z < 0.20) = 0.08$  ... (2)  
and  $f(1.75) = P(0 < Z < 1.75) = 0.46$

Let  $\mu$  and  $\sigma$  be the mean and standard deviation of the distribution. Then  $X \sim N(\mu, \sigma^2)$ .

When  $X = 75$ ,  $Z = \frac{75 - \mu}{\sigma} = z_1$  (say),

and when  $X = 80$ ,  $Z = \frac{80 - \mu}{\sigma} = z_2$  (say).

Thus from the figure, it is obvious that

$$P(X < 75) = 0.58 \Rightarrow P(0 < Z < z_1) = 0.08$$

$\therefore$  Using (2), we have

$$z_1 = \frac{75 - \mu}{\sigma} = 0.20 \quad \dots (3)$$

Also  $P(X > 80) = 0.04 \Rightarrow P(0 < Z < z_2) = 0.46$

$\therefore$  From (2), we get

$$z_2 = \frac{80 - \mu}{\sigma} = 1.75 \quad \dots (4)$$

Solving the equations (3) and (4), we get

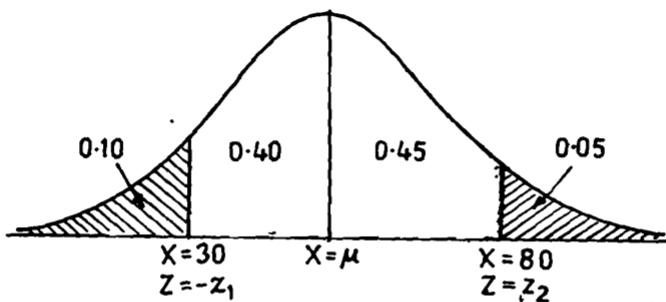
$$\mu = 74.4 \text{ (approx.) and } \sigma = 3.2 \text{ (approx.)}$$

**Example 8.22.** In an examination it is laid down that a student passes if he secures 30 per cent or more marks. He is placed in the first, second or third division according as he secures 60% or more marks, between 45% and 60% marks and marks between 30% and 45% respectively. He gets distinction in case he secures 80% or more marks. It is noticed from the result that 10% of the students failed in the examination, whereas 5% of them obtained distinction. Calculate the percentage of students placed in the second division. (Assume normal distribution of marks.) [Aligarh Univ. B.Sc., 1991]

**Solution.** Let the variable  $X$  denote the marks (out of 100) in the examination and let  $X \sim N(\mu, \sigma^2)$ . Then we are given

$$P(X < 30) = 0.10 \text{ and } P(X \geq 80) = 0.05$$

Thus from the figure on next page, we have



When  $X = 30$ ,  $Z = \frac{30 - \mu}{\sigma} = -z_1$  (say),

and when  $X = 80$ ,  $Z = \frac{80 - \mu}{\sigma} = z_2$  (say).

$$\therefore P(0 < Z < z_2) = 0.5 - 0.05 = 0.45$$

$$\text{and } P(0 < Z < z_1) = P(-z_1 < Z < 0) \\ = 0.50 - 0.10 = 0.40 \quad (\text{By symmetry})$$

$\therefore$  From normal tables, we get

$$z_1 = 1.28 \text{ and } z_2 = 1.64$$

Hence  $\frac{30 - \mu}{\sigma} = -1.28$

$$\Rightarrow \frac{\mu - 30}{\sigma} = 1.28 \text{ and } \frac{80 - \mu}{\sigma} = 1.64$$

Adding, we get

$$\frac{50}{\sigma} = 2.92 \Rightarrow \sigma = \frac{50}{2.92} = 17.12$$

$$\therefore \mu = 30 + 1.28 \times 17.12 = 30 + 21.9136 = 51.9136 \approx 52$$

The probability ' $p$ ' that a candidate is placed in the second division is equal to the probability that his score lies between 45 and 60, i.e.,

$$\begin{aligned} p &= P(45 < X < 60) = P(-0.41 < Z < 0.47) & \left[ Z = \frac{X - 52}{17.12} \right] \\ &= P(-0.41 < Z < 0) + P(0 < Z < 0.47) \\ &= P(0 < Z < 0.41) + P(0 < Z < 0.47) & (\text{By symmetry}) \\ &= 0.1591 + 0.1808 = 0.3399 = 0.34 \text{ (approx.)} \end{aligned}$$

Therefore, 34% candidates got second division in the examination.

**Example 8.23.** The local authorities in a certain city instal 10,000 electric lamps in the streets of the city. If these lamps have an average life of 1,000 hours with a standard deviation of 200 hours, assuming normality, what number of lamps might be expected to fail (i) in the first 800 burning hours? (ii) between 800 and 1,200 burning hours? After what period of burning hours would you expect that (a) 10% of the lamps would fail? (b) 10% of the lamps would be still burning?

[In a normal curve, the area between the ordinates corresponding to  $\frac{X-\bar{X}}{\sigma} = 0$  and  $\frac{X-\bar{X}}{\sigma} = 1$  is 0.34134 and 80% of the area lies between the ordinates corresponding to  $\frac{X-\bar{X}}{\sigma} = \pm 1.28$  ].

**Solution.** If the variable  $X$  denotes the life of a bulb in burning hours, then we are given that  $X \sim N(\mu, \sigma^2)$ , where  $\mu = 1,000$  and  $\sigma = 200$ .

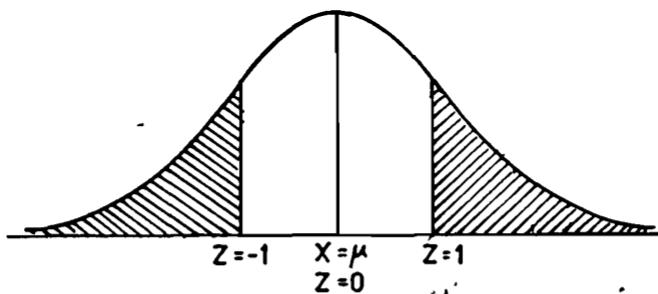
(i) The probability 'p' that bulb fails in the first 800 burning hours is given by

$$\begin{aligned} p &= P(X < 800) = P(Z < -1) = P(Z > 1) \\ &= 0.5 - P(0 < Z < 1) = 0.5 - 0.3413 = 0.1587 \end{aligned} \quad \left[ Z = \frac{800 - 1000}{200} \right]$$

Therefore out of 10,000 bulbs, the number of bulbs which fail in the first 800 hours is

$$10,000 \times 0.1587 = 1587$$

$$\begin{aligned} (ii) \text{ Required probability} &= P(800 < X < 1200) = P(-1 < Z < 1) \\ &= 2P(0 < Z < 1) = 2 \times 0.3413 = 0.6826 \end{aligned}$$



Hence the expected number of bulbs with life between 800 and 1,200 hours of burning life is:  $10,000 \times 0.6826 = 6826$

(a) Let 10% of the bulbs fail after  $x_1$  hours of burning life. Then we have to find  $x_1$  such that  $P(X < x_1) = 0.10$

$$\text{When } X = x_1, \quad Z = \frac{x_1 - 1000}{200} = -z_1 \text{ (say).}$$

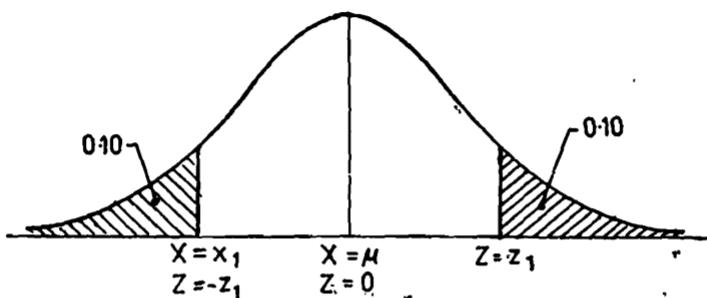
$$\therefore P(Z < -z_1) = 0.10 \Rightarrow P(Z > z_1) = 0.10$$

$$\Rightarrow P(0 < Z < z_1) = 0.40 \quad \dots(1)$$

We are given that

$$P(-1.28 < Z < 1.28) = 0.80 \Rightarrow 2P(0 < Z < 1.28) = 0.80$$

$$\Rightarrow P(0 < Z < 1.28) = 0.40 \quad \dots(2)$$



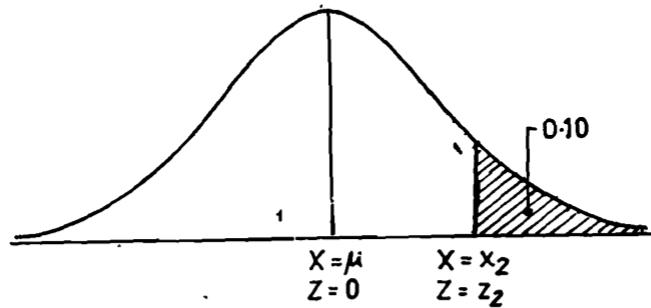
∴ From (1) and (2), we get

$$z_1 = 1.28$$

$$\text{Hence } \frac{x_1 - 1000}{200} = -1.28 \Rightarrow x_1 = 1000 - 256 = 744$$

Thus after 744 hours of burning life, 10% of the bulbs will fail.

(b) Let 10% of the bulbs be still burning after, (say),  $x_2$  hours of burning life. Then we have



$$P(X > x_2) = 0.10 \Rightarrow P(Z > z_2) = 0.10$$

$$\left[ z_2 = \frac{x_2 - 1000}{200} \right]$$

$$\Rightarrow P(0 < Z < z_2) = 0.40$$

$$\therefore z_2 = 1.28$$

[From (2)]

$$\text{i.e., } \frac{x_2 - 1000}{200} = 1.28 \Rightarrow x_2 = 1000 + 256 = 1256$$

Hence after 1256 hours of burning life, 10% of the bulbs will be still burning.

**Example 8-24.** Let  $X \sim N(\mu, \sigma^2)$ . If  $\sigma^2 = \mu^2$ , ( $\mu > 0$ ), express  $P(X < -\mu | X < \mu)$  in terms of cumulative distribution function of  $N(0, 1)$ .

[Delhi Univ. B.Sc. (Maths. Hons.) 1988; (Stat. Hons.). 1993]

**Solution.**

$$P(X < -\mu | X < \mu) = \frac{P(X < -\mu \cap X < \mu)}{P(X < \mu)} = \frac{P(X < -\mu)}{P(X < \mu)}; \quad (\because \mu > 0)$$

$$\begin{aligned}
 &= \frac{P(Z < -2)}{P(Z < 0)} \\
 &= \frac{P(Z > 2)}{(1/2)} ; \quad \text{(By symmetry)}
 \end{aligned}$$

$$= 2 [1 - P(Z \leq 2)] = 2 [1 - \Phi(2)]$$

where  $\Phi(\cdot)$  is the distribution function of standard normal variate.

**Example 8.25** Can  $X$  and  $-X$  have the same distribution?

If so, when? [Delhi Univ. B.A., (Spl. Course Statistics), 1989]

**Solution.** Yes;  $X$  and  $-X$  can have the same distribution provided the p.d.f.  $f(x)$  of  $X$  is symmetric about origin i.e., if  $f(-x) = f(x)$ .

For example,  $X$  and  $-X$  have the same distribution if:

(i)  $X \sim N(0, 1)$

(ii)  $X$  has standard Cauchy distribution [c.f. § 8.9]

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{(1+x^2)} ; \quad -\infty < x < \infty$$

(iii)  $X$  has standard Laplace distribution [c.f. § 8.7]

$$p(x) = \frac{1}{2} e^{-|x|} ; \quad -\infty < x < \infty .$$

and so on. Obviously  $X$  and  $Y = -X$  are not identical.

**Remark.** This example illustrates that if the r.v.'s.  $X$  and  $Y$  are identical, they have the same distributions. However if  $X$  and  $Y$  have the same distribution, it does not imply that they are identical.

**Example 8.26.** If  $X, Y$  are independent normal variables with means 6, 7 and variances 9, 16 respectively, determine  $\lambda$  such that

$$P(2X + Y \leq \lambda) = P(4X - 3Y \geq 4\lambda)$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1988; B.Sc., 1987]

**Solution.** Since  $X$  and  $Y$  are independent, by § 8.2.8 [c.f. equation (8.15a)], we have

$$U = 2X + Y \sim N(2 \times 6 + 7, 4 \times 9 + 16), \text{ i.e., } U \sim N(19, 52)$$

$$V = 4X - 3Y \sim N(4 \times 6 - 3 \times 7, 16 \times 9 + 9 \times 16), \text{ i.e., } V \sim N(3, 288)$$

and  $P(2X + Y \leq \lambda) = P(U \leq \lambda) = P\left(Z \leq \frac{\lambda - 19}{\sqrt{52}}\right)$ , where  $Z \sim N(0, 1)$

and  $P(4X - 3Y \geq 4\lambda) = P(V \geq 4\lambda) = P\left(Z \geq \frac{4\lambda - 3}{12\sqrt{2}}\right)$ , where  $Z \sim N(0, 1)$

Now  $P(2X + Y \leq \lambda) = P\{(4X - 3Y) \geq 4\lambda\}$

$$\Rightarrow P\left(Z \leq \frac{\lambda - 19}{\sqrt{52}}\right) = P\left(Z \geq \frac{4\lambda - 3}{12\sqrt{2}}\right)$$

$$\Rightarrow \frac{\lambda - 19}{\sqrt{52}} = -\frac{4\lambda - 3}{12\sqrt{2}} \quad \therefore$$

[Since  $P(Z \leq a) = P(Z \geq b) \Rightarrow a = -b$ ,

because normal probability curve is symmetric about  $Z = 0$ ].

$$\begin{aligned} \Rightarrow & \frac{\lambda - 19}{\sqrt{13}} = \frac{3 - 4\lambda}{6\sqrt{2}} \\ \Rightarrow & (6\sqrt{2} + 4\sqrt{13})\lambda = 114\sqrt{2} + 3\sqrt{13} \\ \Rightarrow & \lambda = \frac{114\sqrt{2} + 3\sqrt{13}}{6\sqrt{2} + 4\sqrt{13}} \end{aligned}$$

**Example 8.27.** If  $X$  and  $Y$  are independent normal variates possessing a common mean  $\mu$  such that

$$\begin{aligned} P(2X + 4Y \leq 10) + P(3X + Y \leq 9) &= 1 \\ P(2X - 4Y \leq 6) + P(Y - 3X \geq 1) &= 1, \end{aligned}$$

determine the values of  $\mu$  and the ratio of the variances of  $X$  and  $Y$ .

**Solution.** Let  $\text{Var}(X_1) = \sigma_1^2$  and  $\text{Var}(Y) = \sigma_2^2$

Since  $E(X) = E(Y) = \mu$ , (Given) and  $X$  and  $Y$  are independent by § 8.2.8 [c.f. equation (8.15a)], we have

$$2X + 4Y \sim N(2\mu + 4\mu, 4\sigma_1^2 + 16\sigma_2^2), \text{ i.e., } N(6\mu, 4\sigma_1^2 + 16\sigma_2^2)$$

$$3X + Y \sim N(3\mu + \mu, 9\sigma_1^2 + \sigma_2^2), \text{ i.e., } N(4\mu, 9\sigma_1^2 + \sigma_2^2)$$

$$2X - 4Y \sim N(2\mu - 4\mu, 4\sigma_1^2 + 16\sigma_2^2), \text{ i.e., } N(-2\mu, 4\sigma_1^2 + 16\sigma_2^2)$$

$$Y - 3X \sim N(\mu - 3\mu, \sigma_2^2 + 9\sigma_1^2), \text{ i.e., } N(-2\mu, 9\sigma_1^2 + \sigma_2^2)$$

Let us further write :

$$4\sigma_1^2 + 16\sigma_2^2 = \alpha^2 \quad \text{and} \quad 9\sigma_1^2 + \sigma_2^2 = \beta^2 \quad \dots(1)$$

If  $Z$  denotes the Standard Normal Variate, i.e., if  $Z \sim N(0, 1)$ , we get

$$\begin{aligned} & P(2X + 4Y \leq 10) + P(3X + Y \leq 9) = 1 \\ \Rightarrow & P\left(Z \leq \frac{10 - 6\mu}{\alpha}\right) + P\left(Z \leq \frac{9 - 4\mu}{\beta}\right) = 1 \\ \Rightarrow & P\left(Z \leq \frac{10 - 6\mu}{\alpha}\right) = 1 - P\left(Z \leq \frac{9 - 4\mu}{\beta}\right) = P\left(Z \geq \frac{9 - 4\mu}{\beta}\right) \\ \Rightarrow & \frac{10 - 6\mu}{\alpha} = -\left(\frac{9 - 4\mu}{\beta}\right), \end{aligned} \quad \dots(2)$$

(Since normal distribution is symmetric about  $Z = 0$ ).

Similarly

$$\begin{aligned} & P(2X - 4Y \leq 6) + P(Y - 3X \geq 1) = 1 \\ \Rightarrow & P\left(Z \leq \frac{6 + 2\mu}{\alpha}\right) + P\left(Z \geq \frac{1 + 2\mu}{\beta}\right) = 1 \\ \Rightarrow & P\left(Z \leq \frac{6 + 2\mu}{\alpha}\right) = 1 - P\left(Z \geq \frac{1 + 2\mu}{\beta}\right) = P\left(Z \leq \frac{1 + 2\mu}{\beta}\right) \\ \Rightarrow & \frac{6 + 2\mu}{\alpha} = \frac{1 + 2\mu}{\beta} \end{aligned} \quad \dots(3)$$

Solving (2) and (3), we get

$$\frac{\alpha}{\beta} = \frac{6 + 2\mu}{1 + 2\mu} = \frac{10 - 6\mu}{4\mu - 9} \quad \dots(4)$$

$$\Rightarrow (6 + 2\mu)(4\mu - 9) = (10 - 6\mu)(1 + 2\mu)$$

$$\Rightarrow 5\mu^2 - 2\mu - 16 = 0 \quad (\text{On simplification})$$

$$\Rightarrow \mu = \frac{2 \pm \sqrt{4 + 320}}{10} = \frac{2 \pm 18}{10}$$

$$\Rightarrow \mu = 2 \text{ or } -1.6$$

Substituting  $\mu = 2$  in (4), we get.,

$$\frac{\alpha}{\beta} = \frac{10}{5} = 2, \text{ i.e., } 4 = \frac{\alpha^2}{\beta^2}$$

From (1), we get

$$4 = \frac{4\sigma_1^2 + 16\sigma_2^2}{9\sigma_1^2 + \sigma_2^2} = \frac{4 + 16\lambda}{9 + \lambda}$$

$$\Rightarrow 4(9 + \lambda) = 4 + 16\lambda \Rightarrow \lambda = \frac{32}{12} = \frac{8}{3}$$

Taking  $\lambda = \frac{\sigma_2^2}{\sigma_1^2}$

Again putting  $\mu = -1.6$  in (4), we get

$$\left(\frac{14}{11}\right)^2 = \frac{\alpha^2}{\beta^2} = \frac{4 + 16\lambda}{9 + \lambda} \Rightarrow \lambda = \frac{1280}{1740} = \frac{64}{87}$$

**Example 8.28.** If two normal universes A and B have the same total frequency but the standard deviation of universe A is k times that of the universe B, show that maximum frequency of universe A is  $l/k$  times that of universe B.

**Solution.** Let  $N$  be the same total frequency for each of the two universes A and B. If  $\sigma$  is the standard deviation of universe B, then the standard deviation of universe A is  $k\sigma$ . Let  $\mu_1$  and  $\mu_2$  be the means of the universes A and B respectively.

The frequency function of universe A is given by

$$f_A(x) = \frac{N}{k\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_1)^2}{2k^2\sigma^2}\right\}$$

and the frequency function of universe B is given by

$$f_B(x) = \frac{N}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma^2}\right\}$$

Since, for a normal distribution, the maximum frequency occurs at the point  $x = \text{mean}$ , we have

$$[f_A(x)]_{\max} = \text{Maximum frequency of universe A}$$

$$= \left[ f_A(x) \right]_{x=\mu_1}$$

$$= \left[ \frac{N}{k\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_1)^2}{2k^2\sigma^2}\right\} \right]_{x=\mu_1} = \frac{N}{k\sigma\sqrt{2\pi}}$$

Similarly

$$[f_B(x)]_{\max} = [f_B(x)]_{x=\mu_2}$$

$$= \left[ \frac{N}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{(x-\mu_2)^2}{2\sigma^2} \right\} \right]_{x=\mu_1} = \frac{N}{\sigma \sqrt{2\pi}}$$

$$\therefore \frac{|\hat{f}_A(x)|_{\max}}{|\hat{f}_B(x)|_{\max}} = \frac{1}{k}$$

### EXERCISE 8 (b)

1. "If the Poisson and the Normal distributions are limiting cases of Binomial distribution, then there must be a limiting relation between the Poisson and the Normal distributions." Investigate the relation.

2. (a) Derive the mathematical form and properties of normal distribution Discuss the importance of normal distribution in Statistics.

(b) Mention the chief characteristics of Normal distribution and Normal probability curve. [Delhi Univ. B.Sc. (Stat Hons.), 1989]

3. (a) Explain, under what conditions and how the binomial distribution can be approximated to the normal distribution.

(b) For a normal distribution with mean ' $\mu$ ' and standard deviation  $\sigma$ , show that the mean deviation from the mean ' $\mu$ ' is equal to  $\sigma \sqrt{2/\pi}$ . What will be the mean deviation from median?

(c) The distribution of a variable  $X$  is given by the law:

$$f(x) = \text{Constant} \exp \left[ -\frac{1}{2} \left( \frac{x-100}{5} \right)^2 \right], -\infty < x < \infty$$

Write down the value of :

(i) the constant,

(v) standard deviation,

(ii) the mean,

(vi) the mean deviation.

(iii) the median.

(vii) the quartile deviation of the distribution.

(iv) the mode,

[Gujarat Univ. B.Sc. April 1978]

**Ans.** (i)  $\frac{1}{5\sqrt{2\pi}}$ , (ii) 100, (iii) 100, (iv) 100, (v) 5 (vi)  $\sqrt{2/\pi} \times 5 \approx 4$ ,

(vii)  $\frac{1}{2} \times 5 = 3.33$  (approx.) .

(d) Define Normal probability distribution. If the mean of a Normal population is  $\mu$  and its variance  $\sigma^2$ , what are its (i) mode, (ii) Median, (iii)  $\beta_1$  and  $\beta_2$ ?

(e) For a normal distribution  $N(\mu, \sigma^2)$ :

(i) Show that the mean, the median and the mode coincide.

(ii) Find the recurrence relation between  $\mu_{2n}$  and  $\mu_{2n-2}$ .

(iii) State and prove additive property of normal variates.

(iv) Obtain the points of inflexion for the normal distribution  $N(\mu, \sigma^2)$ .

(v) Obtain mean deviation about mean.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

(f) Show that any linear combination of  $n$  independent normal variates is also a normal variate. [Delhi Univ. B.Sc. (Stat. Hons.), 1989]

(g) Show that for the normal curve :

(i) The maximum occurs at the mean of the distribution, and

(ii) the points of inflexion lie at a distance of  $\pm \sigma$  from the mean, where  $\sigma$  is the standard deviation. [Delhi Univ. M.A. (Eco.), 1987]

(h) Describe the steps involved in fitting a normal distribution to the given data and computing the expected frequencies.

(i) Explain how the normal probability integral

$$\int_0^z \varphi(z) dz,$$

is used in computing normal probabilities.

4. Write a note on the salient features of a normal distribution.  $N(\mu, \sigma^2)$  denotes the normal distribution of each of the random variables  $X_1, X_2, X_3, \dots, X_n$ , where  $\mu$  is the mean and  $\sigma^2$  the variance. Prove the following :

(i) If  $X_1, X_2, \dots, X_n$  are independent, then  $X_1 + X_2 + \dots + X_n$  has the distribution  $N(n\mu, n\sigma^2)$ .

(ii)  $kX$ , where  $k$  is a constant has the distribution  $N(k\mu, k^2\sigma^2)$ .

(iii)  $X+a$ , where  $a$  is a constant has the distribution  $N(\mu+a, \sigma^2)$

(iv) In (i) if  $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$  then

$\sqrt{n} \frac{(\bar{X} - \mu)}{\sigma}$  has the distribution  $N(0, 1)$ .

5. (a) Show that for a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , the central moments satisfy the relation

$$\mu_{2n} = (2n-1) \mu_{2n-2} \sigma^2; \mu_{2n+1} = 0$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

Hence show that  $\mu_{2n} = \frac{(2n)!}{n!} (\frac{1}{2} \sigma^2)^n$  and  $\mu_{2n+1} = 0$ ;  $n = 1, 2, \dots$

[Delhi Univ. B.Sc. (Stat Hons.) 1985]

(b) State the mathematical equation of a normal curve. Discuss its chief features.

(c) Find the moment generating function of the normal distribution  $(m, \sigma^2)$ , and deduce that

$$\mu_{2n+1} = 0,$$

$$\mu_{2n} = 1 \cdot 3 \cdot 5 \dots (2n-1) \sigma^{2n},$$

where  $\mu_n$  denotes the  $n$ th central moment.

[Delhi Univ. B.Sc. (Stat. Hons.) 1990, '82]

(d) Show that all central moments of a normal distribution can be expressed in terms of the standard deviation and obtain the expression in the general case.

[Aligarh Univ. B.Sc. 1992]

(e) The normal table gives the values of the integral:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{1}{2}t^2\right) dt$$

for different values of  $x$ .

Explain how to use this table to obtain the proportion of observations of a normal variate with mean  $\mu$  and S.D.  $\sigma$ , which lie above a given value ' $a$ ',

(i) where  $a > \mu$ , (ii) where  $a < \mu$ .

6. (a) If  $X_1$  and  $X_2$  are two independent normal variates with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$  respectively, show that the variables  $U$  and  $V$  where  $U = X_1 + X_2$  and  $V = X_1 - X_2$ , are independent normal variates. Find the means and variances of  $U$  and  $V$ .

(b) If  $X_1$  and  $X_2$  are independent standard normal variates obtain the p.d.f. of  $(X_1 - X_2)/\sqrt{2}$ .

$$\text{Ans. } U = (X_1 - X_2)/\sqrt{2} \sim N(0, 1)$$

(c) Suppose  $X_1 \sim N(0, 1)$  and  $X_2 \sim N(0, 1)$  are independent r.v.'s.

(i) Find the joint distribution of  $(X_1 + X_2)/\sqrt{2}$  and  $(X_1 - X_2)/\sqrt{2}$ .

(ii) Argue that  $2X_1 X_2$  and  $X_2^2 - X_1^2$  have the same distribution.

Ans. (i)  $U = (X_1 + X_2)/\sqrt{2}$  and  $V = (X_1 - X_2)/\sqrt{2}$  are independent  $N(0, 1)$  variates

$$(ii) \text{ Hint. } X_2^2 - X_1^2 = 2 \left[ \frac{X_2 + X_1}{\sqrt{2}} \right] \left[ \frac{X_2 - X_1}{\sqrt{2}} \right] = 2(UV)$$

$$= 2 \times [\text{Product of two independent SNV's}]$$

$$2X_1 X_2 = 2 \times [\text{Product of two independent SNV's}]$$

Hence the result.

7. (a) Let  $X$  be normally distributed with mean 8 and s.d. 4. Find

(i)  $P(5 \leq X \leq 10)$ , (ii)  $P(10 \leq X \leq 15)$ , (iii)  $P(X \geq 15)$ , (iv)  $P(X \leq 5)$ .

Ans. (i) 0.4649 (ii) 0.2684 (iii) 0.0401 (vi) 0.2266.

(b) The standard deviation of a certain group of 1,000 high school grades was 11% and the mean grade 78%. Assuming the distribution to be normal, find

(i) How many grades were above 90%?

(ii) What was the highest grade of the lowest 10%?

(iii) What was the interquartile range?

(iv) Within what limits did the middle 90% lie?

Ans. (i) 138, (ii) 52, (iii)  $Q_1 = 70.575$ ;  $Q_3 = 85.425$ , and (iv) 60% to 96.2%

(c) If  $X$  is normally distributed with mean 2 and variance 1, find

$$P(|X - 2| < 1).$$

$$\text{Ans. } 0.6826 \text{ [or } \Phi(1) - \Phi(-1)]$$

(d) If  $X \sim N(\mu = 2, \sigma^2 = 2)$ , find  $P(|X - 1| \leq 2)$  in terms of distribution function of standard normal variate.

**Ans.** Probability  $= P(-1 \leq X \leq 3) = \Phi(1/\sqrt{2}) - \Phi(-3/\sqrt{2})$

(e) If  $X \sim N(30, 5^2)$  and  $Y \sim N(15, 10^2)$ , show that

$$P(26 \leq X \leq 40) = P(7 \leq Y \leq 35).$$

**Hint.** Each Probability  $= P(-0.8 \leq Z \leq 2)$  where  $Z \sim N(0, 1)$

(f) If  $X \sim N(30, 5^2)$ , find the probabilities of

- (i)  $26 \leq X \leq 40$ , (ii)  $|X - 30| > 5$ , (iii)  $X \geq 42$ , (iv)  $X \leq 28$

[Bihar P.C.S., 1988]

**Ans.** (i) 0.7653, (ii) 0.3174, (iii) 0.0082, (iv) 0.3446

8. (a) In a normal population with mean 15.00 and standard deviation 3.5, it is known that 647 observations exceed 16.25. What is the total number of observations in the population? (Sri Venkateswara Univ. B.Sc. April 1990)

**Hint.** Let  $X \sim N(\mu, \sigma^2)$  where  $\mu = 15$  and  $\sigma = 3.5$ .

If  $N$  is the total number of observations in the population, then we have to find  $N$  such that

$$N \times P(X > 16.25) = 647$$

(b) Assume the mean heights of soldiers to be 68.22 inches with a variance of  $10.8$  (in.) $^2$ . How many soldiers in a regiment of 1,000 would you expect to be over 6 feet tall? (Given that the area under the standard normal curve between  $X = 0$  and  $X = 0.35$  is 0.1368 and between  $X = 0$  and  $X = 1.15$  is 0.3746).

**Ans.** 125

[Osmania Univ. M.A., 1992]

9. (a) If 100 true coins are thrown, how would you obtain an approximation for the probability of getting (i) 55 heads, (ii) 55 or more heads, using Tables of Area of normal probability function.

(b) Prove that Binomial distribution in certain cases becomes normal.

A six faced dice is thrown 720 times. Explain how an approximate value of the probability of the following events can be found out easily. (Finding out the numerical values of these probabilities is not necessary):

(i) 'six' comes for more than 130 times

(ii) chance of 'six' lies between 100 and 140.

10. (a) The number ( $X$ ) of items of a certain kind demanded by customers follows the Poisson law with parameter 9. What stock of this item should a retailer keep in order to have a probability of 0.99 of meeting all demands made on him? Use normal approximation to the Poisson law.

(b) Show that the probability that the number of heads in 400 throws of a fair coin lies between 180 and 220 is  $\approx 2F(2) - 1$ , where  $F(x)$  denotes the standard normal distribution function.

11. In an intelligence test administered to 1,000 children, the average score is 42 and standard deviation 24.

(i) Find the number of children exceeding the score 60, and

(ii) Find the number of children with score lying between 20 and 40.  
(Assume the normal distribution.) Ans. (i) 227 (iii) 289

12. The mean I.Q. (intelligence quotient) of a large number of children of age 14 was 100 and the standard deviation 16. Assuming that the distribution was normal, find

(i) What % of the children had I.Q. under 80?

(ii) Between what limits the I.Q.'s of the middle 40% of the children lay?

(iii) What % of the children had I.Q.'s within the range  $\mu \pm 1.96\sigma$ ?

Ans. (i) 10.56%, (ii) 91.6, 108.4, (iii) 0.95

13. (a) In a university examination of a particular year, 60% of the students failed when mean of the marks was 50% and s.d. 5%. University decided to relax the conditions of passing by lowering the pass marks, to show its result 70%. Find the minimum marks for a student to pass, supposing the marks to be normally distributed and no change in the performance of students takes place.

Ans. 47.375.

(b) The width of a slot on a forging is normally distributed with mean 0.900 inch and standard deviation 0.004 inch. The specifications are  $0.900 \pm 0.005$  inch. What percentage of forgings will be defective?

**Hint.** Let  $X$  denote the width (in inches) of the slot. We want

$$100 \times P(X \text{ lies outside specification limits})$$

$$= 100 [1 - P(X \text{ lies within specification limits})]$$

$$= 100 [1 - P(0.895 < X < 0.905)].$$

14. (a) The monthly incomes of a group of 10,000 persons were found to be normally distributed with mean Rs. 750 and s.d. Rs. 50. Show that of this group, about 95% had income exceeding Rs. 668 and only 5% had income exceeding Rs. 832. What was the lowest income among the richest 100?

Ans. Rs. 866.3.

(b) Given that  $X$  is normally distributed with mean 10 and

$$P(X > 12) = 0.1587,$$

what is the probability that  $X$  will fall in the interval (9, 11)?

Take  $\Phi(1) = 0.8413$  and  $\Phi(-\frac{1}{2}) = 0.3085$

$$\text{where } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-u^2/2) du$$

Ans. 0.3830

(c) A normal distribution has mean 25 and variance 25. Find

(i) the limits which include the middle 50% of the area under the curve, and  
(ii) the values of  $x$  corresponding to the points of inflection of the curve.

Ans. (i) Limits which include the middle 50% of the area under the curve are:

$$Q_1 = \mu - 0.6745\sigma = 21.7275; Q_3 = \mu + 0.6745\sigma = 38.2725$$

$$(ii) (30, 20)$$

15. (a) In a distribution exactly normal 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution?

Ans.  $\mu = 50.3$ ,  $\sigma = 10.33$ .

(b) In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and variance of the distribution.

Given that area between mean-ordinates and ordinate at any  $\sigma$  distance from mean,

$$Z = \frac{X - \mu}{\sigma} : 0.496 \quad 1.405$$

$$\text{Area} : 0.19 \quad 0.42$$

[Delhi Univ. B.Sc., 1987; Madras Univ. B.Sc., 1990]

Ans.  $\mu = 50$ ,  $\sigma = 10$

16. (a) A minimum height is to be prescribed for eligibility to government services such that 60% of the young men will have a fair chance of coming up to that standard. The heights of youngmen are normally distributed with mean 60.6" and s.d. 2.55". Determine the minimum specification.

Ans. 59.9".

**Hint.** We want  $x_1$  s.t.  $P(X > x_1) = 0.6$

$$\text{When } X = x_1, Z = \frac{x_1 - 60.6}{2.55} = -z_1, (\text{say}) \dots (*)$$

[Note the negative sign, which is obvious from the diagram]

$$\text{Obviously } P(0 < Z < z_1) = 0.10 \Rightarrow z_1 = 0.254$$

Substituting in (\*), we get

$$x_1 = 60.6 - 2.55 \times 0.254 = 60.6 - 0.65 = 59.95"$$

(b) The height measurements of 600 adult males are arranged in ascending order and it is observed that 180th and 450th entries are 64.2" and 67.8" respectively. Assuming that the sample of heights is drawn from a normal population, estimate the mean and s.d. of the distribution.

Ans. 67.78", 3"

17. (a) Marks secured by students in sections I and II of a class are independently normally distributed with means 50 and 60 respectively and variances 10 and 6 respectively. What is the probability that a randomly chosen student from section II scores more marks than a randomly chosen student from section I? What percentage of students are expected to secure first division (i.e., 60 marks or more) in section I? Write down your results in terms of the standard normal distribution function.

**Hint.**  $X \sim N(50, 10)$ ,  $Y \sim N(60, 6)$  are independent r.v.'s.

$$U = Y - X \sim N(10, 16). \text{ We want } P(Y > X) = P(U > 0).$$

(b) In an examination, the mean and standard deviation (s.d.) of marks in Mathematics and Chemistry are given below

	Mean	s.d.
Maths.	45	10
Chem.	50	15

Assuming the marks in the two subjects to be independent normal variates, obtain the probability that a student scores total marks lying between 100 and 130. [Full marks in each subject are 100]. Given that

$$F(0.28) = 0.1103, F(1.94) = 0.4738,$$

where

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_0^z \exp(-\frac{1}{2}x^2) dx.$$

[Bhagalpur Univ. B.Sc., 1990]

18. (a) One thousand candidates in an examination were grouped into three classes I, II, III in descending order of merits. The numbers in the first two classes were 50 and 350 respectively. The highest and the lowest marks in class II were 60 and 50 respectively. Assuming the distribution to be normal, prove that the average mark is approximately 48.2 and standard deviation, approximately 7.1. The following data may be used:

The area  $A$  is measured from the mean zero to any ordinate  $X$ .

$\frac{X}{\sigma}$	$A$	$\frac{X}{\sigma}$	$A$
0.2	0.079	1.5	0.433
0.3	0.118	1.6	0.445
0.4	0.155	1.7	0.455

(b) In an examination marks obtained by the students in Mathematics, Physics and Chemistry are distributed normally about the means 50, 52 and 48 with S.D. 15, 12, 16 respectively. Find the probability of securing total marks of

- (i) 180 or above, (ii) 90 or below.

$$\left[ \frac{1}{\sqrt{2\pi}} \int_{-1.2}^{\infty} \exp(-z^2/2) dz = 0.1942, \quad \frac{1}{\sqrt{2\pi}} \int_{-2.4}^{\infty} \exp(-z^2/2) dz = 0.0224 \right]$$

Ans. 0.1942, 0.0224

[Agra Univ. B.Sc., 1988]

19. In a certain examination the percentage of passes and distinctions were 46 and 9 respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75 respectively. (Assume the distribution of marks to be normal.) (Ans.  $\mu = 36.4$ ,  $\sigma = 28.2$ )

Also determine what would have been the minimum qualifying marks for admission to a re-examination of the failed candidates, had it been desired that the best 25% of them should be given another opportunity of being examined.

Ans. 29.

20. The local authorities in a certain city installed 2,000 electric lamps in a street of the city. If the lamps have an average life of 1,000 burning hours with a S.D. of 200 hours,

(i) What number of the lamps might be expected to fail in the first 700 burning hours,

(ii) After what periods of burning hours would we expect that

(a) 10% of the lamps would have failed, and

(b) 10% of the lamps would be still burning?

Assume that lives of the lamps are normally distributed.

You are given that  $F(1.50) = 0.933$ ,  $F(1.28) = 0.900$ ,

where 
$$F(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Ans. (i) 134, (ii) (a) 744, (b) 1256. [Allahabad Univ. B.Sc., 1987]

21. (a) The quartiles of a normal distribution are 8 and 14 respectively. Estimate the mean and standard deviation.

Ans.  $\mu = 11$ ,  $\sigma = 4.4$ .

(b) The third decile and the upper quartile of a normal distribution are 56 and 63 respectively. Find the mean and variance of the distribution.

Ans.  $\mu = 59.1$ ,  $\sigma = 5.8$ .

22. (a) 5,000 variates are normally distributed with mean 50 and probable error (semi-interquartile range) 13.49. Without using tables, find the values of the quartiles, median, mode standard deviation and mean deviation. Find also the value of the variate for which cumulative frequency is 1250.

[Meerut Univ. B.Sc., 1989]

Ans.  $Q_1 = 36.51$ ,  $Q_3 = 63.49$ ,  $\sigma = 20$ , M.D. = 16,  $x_1 = 36.51$ .

(b) The following table gives frequencies of occurrence of a variable X between certain limits :

Variable X	Frequency
Less than 40	30
40 or more but less than 50	33
50 and more	37

The distribution is exactly normal. Find the distribution and also obtain the frequency between  $X = 50$  and  $X = 60$ . [Kurukshetra Univ. M.A. (Eco.), 1990]

Ans. Hint.  $50 - \mu = 0.33 \sigma$ ;  $40 - \mu = -0.52 \sigma$

$$\mu = 46.12, \sigma = 11.76$$

$$N.P(50 < X < 60) = 100 \times 0.2517 \approx 25$$

23. (a) Suppose that a doorway being constructed is to be used by a class of people whose heights are normally distributed with mean 70" and standard deviation 3". How long may the doorway be without causing more than 25% of the

people to bump their heads? If the height of the doorway is fixed at 76", how many persons out of 5,000 are expected to bump their heads?

[For a normal distribution the quartile deviation is 0.6745 times standard deviation. For a standard normal distribution  $Z = \frac{X - \bar{X}}{\sigma}$ , the area under the curve between  $Z = 0$  and  $Z = 2$  is 0.4762.]

(b) A normal population has a coefficient of variation 2% and 8% of the population lies above 120. Find the mean and S.D.

**Ans.**  $\mu = 122$ ,  $\sigma = 2.44$

24. Steel rods are manufactured to be 3 inches in diameter but they are acceptable if they are inside the limits 2.99 inches and 3.01 inches. It is observed that 5% are rejected as oversize and 5% are rejected as undersize. Assuming that the diameters are normally distributed, find the standard deviation of the distribution. Hence calculate, what would be the proportion of rejects if the permissible limits were widened to 2.985 inches and 3.015 inches.

[Hint. Let  $X$  denote the diameter of the rods in inches and let  $X \sim N(\mu, \sigma^2)$ .

Then we are given

$$\begin{aligned} P(X > 3.01) &= 0.05 \text{ and } P(X < 2.99) = 0.05 \\ \Rightarrow \frac{3.01 - \mu}{\sigma} &= 1.65 \text{ and } \frac{2.99 - \mu}{\sigma} = -1.65 \end{aligned}$$

$$\text{Solving we get } \mu = 3 \text{ and } \sigma = \frac{1}{165}$$

$$\begin{aligned} \text{The probability that a random value of } X \text{ lies within the rejection limits is} \\ P(2.985 < X < 3.015) &= P(-2.475 < Z < 2.475) = 2 \times P(0 < Z < 2.475) \\ &= 2 \times 0.4933 = 0.9866 \end{aligned}$$

Hence the probability that  $X$  lies outside the rejection limits is

$$1 - 0.9866 = 0.0134$$

Therefore, the proportion of the rejects outside the revised limits is 0.0134, i.e., 1.34%].

25. Derive the moment generating function of a random variable which has a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Hence or otherwise prove that a linear combination of independent normal variates is also normally distributed.

An investor has the choice of two of four investments  $X_1, X_2, X_3, X_4$ . The profits from these may be assumed to be independently distributed, and

the profit from  $X_1$  is  $N(2, 1)$ ,

the profit from  $X_2$  is  $N(3, 3)$ ,

the profit from  $X_3$  is  $N(1, \frac{1}{4})$ .

the profit from  $X_4$  is  $N(2 \frac{1}{2}, 4)$ .

(Profits are given in £ 1000 per annum).

Which pair should he choose to maximise his probability of making a total annual profit of at least £ 2000? (London Univ. B.Sc. 1977)

26. (a) State the important properties of the normal distribution and obtain from the tables the inter-quartile range in terms of its mean  $\mu$  and standard deviation  $\sigma$ .

Find the mean and standard deviation as well as the inter-quartile range of the following data. Compare the inter-quartile range with that obtained from mean and standard deviation on the assumption of normality.

$X$ (central values) ...	0	1	2	3	4	5	6
$f$ (frequency) ...	5	9	15	32	21	10	8

(b) The following table gives Baseball throws for a distance by 303 first year high school girls:

Distance in feet	Number of girls	Distance in feet	Number of girls
15 and under 25	1	85 and under 95	44
25 and under 35	2	95 and under 105	31
35 and under 45	7	105 and under 115	27
45 and under 55	25	115 and under 125	11
55 and under 65	33	125 and under 135	4
65 and under 75	53	135 and under 145	1
75 and under 85	64		

(i) Fit a normal distribution and find the theoretical frequencies for the classes of the above frequency distribution.

(ii) Find the expected number of girls throwing baseballs at a distance exceeding 105 feet on the basis that the data fit a normal distribution.

27. (a) The table given below shows the distribution of heights among freshmen in a college :

Height in inches	61	62	63	64	65	66	67	68
Frequency	4	20	23	75	114	186	212	252
Height in inches	69	70	71	72	73	74		
Frequency	218	175	149	46	18	8		

By comparing the proportion of cases lying between  $\mu \pm (2/3)\sigma$ ,  $\mu \pm \sigma$ ,  $\mu \pm 2\sigma$  and  $\mu \pm 3\sigma$ , for this distribution and for a normal curve, state whether the distribution may be considered normal.

(b) Fit a normal distribution to the following data of heights in cms of 200 Indian adult males :

Height in (cms)	Frequency
144 — 150	3
150 — 156	12
156 — 162	23
162 — 168	52
168 — 174	61
174 — 180	39
180 — 186	10

(c) Two hundred and fifty-five metal rods were cut roughly six inches over size. Finally the lengths of the oversize amount were measured exactly and grouped with 1-inch intervals, there being in all 12 groups. The frequency distribution for the 255 lengths was

Central value : $x$	1	2	3	4	5	6
Frequency : $f$	2	10	19	25	40	44
$x$	7	8	9	10	11	12
$f$	41	28	25	15	5	1

Fit a normal distribution to the data by the method of ordinates and calculate the expected frequencies.

28. (a) Let  $X \sim N(\mu, \sigma^2)$ . Let

$$\Phi(x) = P[X \leq x],$$

calculate the probabilities of the following events in terms of  $\Phi$ :

(i)  $\alpha X + \beta \leq t$ , where  $\alpha, \beta$  are finite constants.

(ii)  $-X \geq t$

(iii)  $|X| > t$

[Poona Univ. B.E., 1991]

(b) Determine  $C$  such that the following function becomes a distribution function:

$$F(x) = C \int_{-\infty}^x \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] dy$$

29. (a) Determine the constant  $C$  so that  $C e^{-2x^2+x}$ ,  $-\infty < x < \infty$ , is a density function. If the random variable  $X$  has the resulting density function, then find (i) the mean of  $X$ , (ii) the variance of  $X$  and (iii)  $P(X \geq 1/4)$ .

Ans. (i) 0.25 (ii) 0.25 (iii) 0.5

(b) If  $f(x) = k \exp\{-9x^2 - 12x + 13\}$ , is the p.d.f. of a normal distribution ( $k$ , being a constant) find the mean and s.d. of the distribution.

(c) If  $X$  is a normal variate with p.d.f.  $f(x) = 0.03989 \exp(-0.005x^2 + 0.5x - 12.5)$ , express  $f(x)$  in standard form and hence find the mean and variance of  $X$ . [M.S. Baroda Univ. B.Sc., 1991]

(d) Let the probability function of the normal distribution be

$$P(x) = ke^{-\frac{1}{8}x^2 + 2x}, -\infty < x < \infty$$

Find  $k$ ,  $\mu$  and  $\sigma^2$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1985]

(e)  $X_1, X_2, X_3, X_4$  is a random sample from a normal distribution with mean 100 and variance 25 and  $\bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$ .

State the distribution, expected value and variance of each of the following:

$$(i) 4\bar{X}, (ii) X_1 - 2X_2 + X_3 - 3X_4,$$

$$(iii) \frac{1}{25} \sum_{i=1}^4 (X_i - 100)^2 \quad [\text{Bangalore Univ. B.Sc., 1989}]$$

**Ans.** (b) Mean =  $2/3$ ,  $\sigma = \frac{1}{3\sqrt{2}}$

30. If  $X$  is a normal variate with mean 50 and s.d. 10, find  $P(Y \leq 3137)$ , where  $Y = X^2 + 1$ ,

$$\left[ \frac{1}{\sqrt{2}\pi} \int_0^{0.6} e^{-x^2/2} dx = 0.2258 \right] \quad [\text{Delhi Univ. B.Sc. (Hons.), 1990}]$$

**Hint.** Required Probability =  $P(X^2 + 1 \leq 3137) = P(-56 \leq X \leq 56)$ .

**Ans.** 0.7258

31. Let  $X$  be normally distributed with mean  $\mu$  and variance  $\sigma^2$ . Suppose  $\sigma^2$  is some function of  $\mu$ , say  $\sigma^2 = h(\mu)$ . Pick  $h(\cdot)$  so that  $P(X \leq 0)$  does not depend on  $\mu$  for  $\mu > 0$ .

**Ans.**  $P(X \leq 0) = P(Z \leq -\mu/\sqrt{h(\mu)}) = P(Z \leq -1)$ ; independent of  $\mu$  if we take  $h(\mu) = \mu^2$ .

32. (a) If  $X$  is a standard normal variate, find  $E|X|$  [Ans.  $\sqrt{2/\pi} \approx 4/4$ ]

(b)  $X$  is a random variable normally distributed with mean zero and variance  $\sigma^2$ . Find  $E|X|$  [Delhi Univ. B.Sc. (Stat. Hons.) 1990]

**Hint.**  $E|X| = \text{Mean Deviation about origin}$

= M.D. about mean ( $\because$  Mean = 0)

$$\text{Ans. } \sqrt{(2/\pi)} \cdot \sigma \approx \frac{4}{5} \sigma$$

32. (a)  $X$  is a normal variate with mean 1 and variance 4,  $Y$  is another normal variate independent of  $X$  with mean 2 and variance 3. What is the distribution of  $X + 2Y$ ? [Punjab Univ. B.Sc. (Hons.) 1993]

**Ans.**  $X + 2Y \sim N(5, 16)$

(b) If  $X$  is a normal variate with mean 1 and S.D. 0.6, obtain  $P[X > 0]$ ,  $P[|X - 1| \geq 0.6]$  and  $P[-1.8 < X < 2.0]$ . What is the distribution of  $4X + 5$ ?

34. (a) Let  $X$  and  $Y$  be two independent random variables each with a distribution which is  $N(0, 1)$ . Find the probability density function of  $U = a_1 X + a_2 Y$ , where  $a_1$  and  $a_2$  are constants.

(b) Show that if  $X_1, X_2$  are mutually independent normal variates having means  $\mu_1, \mu_2$  and standard deviations  $\sigma_1, \sigma_2$  respectively, then  $U = a_1 X_1 + a_2 X_2$  is also normally distributed.

34. (c) If  $X_i$ , ( $i = 1, 2, \dots, n$ ) are independent  $N(\mu_i, \sigma_i^2)$  variates, obtain the distribution of  $\sum_{i=1}^n a_i X_i$

where  $a_i$ , ( $i = 1, 2, \dots, n$ ) are constants. Hence deduce the distributions of :

$$(i) X_1 + X_2$$

$$(ii) X_1 - X_2$$

$$(iii) \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i; \text{ if } X_i \text{'s are i.i.d. } N(\mu, \sigma^2).$$

How do the results in (i) and (ii) compare with those in Poisson distribution and result in (iii) compare with Cauchy distribution ?

[Delhi Univ. B.Sc. (Stat. Hons.), 1991]

**Hint.** For Cauchy distribution, see Remark 4, § 8.9.1.

35. (a) If  $X$  is normal with mean 2 and standard deviation 3, describe the distribution of  $Y = \frac{1}{2}X - 1$ . Explain how you would find  $P(Y \geq \frac{3}{2})$  from the tables.

**Hint.** (a) We are given that  $X \sim N(\mu, \sigma^2)$  where  $\mu = 2, \sigma = 3$ . The distribution of the new variable  $Y = aX + b$  is also normal with

$$\begin{aligned} E(Y) &= E(aX + b) = aE(X) + b = a\mu + b \\ \text{and } \text{Var}(Y) &= \text{Var}(aX + b) = a^2 \text{Var}(X) = a^2 \sigma^2 \end{aligned} \quad \dots (*)$$

Hence  $Y = \frac{1}{2}X - 1 \sim N(\mu_1, \sigma_1^2)$ , where  $\mu_1$  and  $\sigma_1^2$  are given by (\*) with  $a = \frac{1}{2}$  and  $b = -1$ , i.e.,

$$\mu_1 = \frac{1}{2} \cdot 2 - 1 = 0; \sigma_1^2 = \left(\frac{1}{2}\right)^2 \cdot 9 = \frac{9}{4}.$$

Thus  $Y \sim N(\mu_1, \sigma_1^2)$ , where  $\mu_1 = 0, \sigma_1 = \frac{3}{2}$ .

$$P(Y \geq \frac{3}{2}) = P(Z \geq 1) = 0.5 - P(0 < Z < 1) = 0.5 - 0.3413 = 0.1587.$$

(b) If  $X$  and  $Y$  are independent standard normal variables and if  $Z = aX + bY + c$  where  $a, b$  and  $c$  are constants, what will be the distribution of  $Z$ ? What is the mean, median and standard deviation of the distribution of  $Z$ ?

Find  $P(Z \leq 0.1)$  if  $a = 1, b = -1$  and  $c = 0$ . (I.I.T. B. Tech. 1992)

**Hint.**  $Z \sim N(c, a^2 + b^2)$

If  $a = 1, b = -1, c = 0$  then  $Z \sim X - Y \sim N(0, 2)$

$$\therefore P(Z \leq 0.1) = P\left(U \leq \frac{1}{14.142}\right); \quad U = \frac{Z-0}{\sqrt{2}} \sim N(0, 1)$$

36. Let  $X$  be a random variable following normal distribution with mean  $\mu$  and variance  $\sigma^2$  and let  $r$  be a non-negative integer.

If  $\mu'_r = E(X^r)$  and if  $\mu_{2r} = [E(X - \mu)^{2r}]$ , prove that

$$(i) \quad \mu'_{r+2} = 2\mu\mu'_{r+1} + (\sigma^2 - \mu^2)\mu_r + \sigma^3 \frac{d\mu'_r}{d\sigma}$$

$$(ii) \quad \mu_{2r+2} = \sigma^2 \mu_{2r} + \sigma^3 \frac{d\mu_{2r}}{d\sigma} \quad [\text{Madras Univ. B.Sc. (Main), Oct. 1989}]$$

**Hint.** (i)

$$\begin{aligned} \frac{d\mu'_r}{d\sigma} &= - \int_{-\infty}^{\infty} \frac{x^r}{\sqrt{2\pi}\sigma^2} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx \\ &\quad + \int_{-\infty}^{\infty} \frac{x^r(x-\mu)^2}{\sqrt{2\pi}\sigma^4} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx \\ &= -\frac{\mu'_r}{\sigma} + \frac{\mu'_{r+2}}{\sigma^3} - \frac{2\mu\mu'_{r+1}}{\sigma^3} + \frac{\mu^2\mu'_r}{\sigma^3} \end{aligned}$$

$$\begin{aligned} (ii) \quad \frac{d\mu_{2r}}{d\sigma} &= - \int_{-\infty}^{\infty} \frac{(x-\mu)^{2r}}{\sqrt{2\pi}\sigma^2} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx \\ &\quad + \int_{-\infty}^{\infty} \frac{(x-\mu)^{2r+2}}{\sqrt{2\pi}\sigma^4} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = -\frac{\mu_{2r}}{\sigma} + \frac{\mu_{2r+2}}{\sigma^3} \end{aligned}$$

37. Prove that if the independent random variables  $X$  and  $Y$  have the probability densities,

$$\frac{h}{\sqrt{\pi}} e^{-h^2 x^2} \text{ and } \frac{k}{\sqrt{\pi}} e^{-k^2 y^2}, \quad -\infty < (x, y) < \infty$$

then the random variable  $U = X + Y$  has the probability density,

$$\frac{l}{\sqrt{\pi}} \cdot e^{-l^2 u^2}, \quad -\infty < u < \infty$$

where

$$\frac{1}{l^2} = \frac{1}{h^2} + \frac{1}{k^2}$$

$$38. \text{ If } \left[ \sum_{i=1}^n c_i \mu_i \right]^2 = 9 \sum_{i=1}^n c_i^2 \sigma_i^2, \text{ find } P\left(0 \leq Y \leq 2 \sum_{i=1}^n c_i \mu_i\right),$$

where  $Y = \sum_{i=1}^n c_i X_i$ ,  $X_i$  being a normal variate with mean  $\mu_i$  and variance  $\sigma_i^2$ .

**Hint.**

We know  $Y = \sum_{i=1}^n c_i X_i \sim N(\mu, \sigma^2)$ , where  $\mu = \sum_{i=1}^n c_i \mu_i$  and  $\sigma^2 = \sum_{i=1}^n c_i^2 \sigma_i^2$

Since  $\left( \sum_i c_i \mu_i \right)^2 = 9 \left( \sum_i c_i^2 \sigma_i^2 \right)$ , we have  $\mu^2 = 9 \sigma^2$  or  $\frac{\mu}{\sigma} = 3$

If we write  $Z = \frac{Y - \mu}{\sigma}$ , then  $Z \sim N(0, 1)$ .

$$P(0 \leq Y \leq 2 \sum_{i=1}^n c_i \mu_i) = P(0 \leq Y \leq 2\mu) = P(-3 \leq Z \leq 3) = 0.9973$$

39. (a). Find the mean deviation about mean for the normal distribution  $N(\mu, \sigma^2)$ .

(b) If  $X \sim N(\mu, \sigma^2)$ , find the mean and variance of

$$Y = \frac{1}{2} [(x - \mu)/\sigma]^2 \quad [\text{Punjabi Univ. M.A. (Eco.), 1991}]$$

**Ans.**  $E(Y) = 1/2$ ,  $\text{Var}(Y) = 1/2$

**Remark.** Also see Example 8.30, on Gamma distribution.

(c) Derive normal distribution as a limiting case of binomial distribution, clearly stating the conditions involved. [Delhi Univ. B.A. (Stat. Hons.), 1981]

40. If  $f(x)$  is the density function for the normal distribution with mean zero and standard deviation  $\sigma$ , then show that

$$\int_{-\infty}^{+\infty} [f(x)]^2 dx = \frac{1}{2\sigma\sqrt{\pi}}$$

Hence show that if the normal distribution is grouped in intervals with total frequency  $N_1$ , and  $N_2$  is the sum of the squares of the frequencies, an estimate of  $\sigma$  is  $\frac{N_1^2}{2N_2\sqrt{\pi}}$  [Gujarat Univ. B.Sc., 1992]

$$\begin{aligned} \text{Hint. } \int_{-\infty}^{+\infty} [f(x)]^2 dx &= \int_{-\infty}^{+\infty} \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp(-x^2/2\sigma^2) \right\}^2 dx \\ &= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{+\infty} e^{-x^2/\sigma^2} dx = \frac{1}{2\pi\sigma^2} \cdot \frac{\sqrt{\pi}}{(1/\sigma)} \\ &= \frac{1}{2\sigma\sqrt{\pi}} \quad \left( \because \int_{-\infty}^{+\infty} e^{-a^2 x^2} dx = \sqrt{\pi}/a \right) \\ N_2 &= \int_{-\infty}^{+\infty} \{N_1 f(x)\}^2 dx = \frac{N_1^2}{2\sqrt{\pi}\sigma} \end{aligned}$$

41. Obtain the normal distribution as a limiting case of Poisson distribution when the parameter  $\lambda \rightarrow \infty$ .

42. (a) If  $X$  is  $N(0, 1)$ , prove that the p.d.f. of  $|X|$  is

$$h(x) = \begin{cases} \sqrt{2/\pi} \exp(-x^2/2), & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

(b) Let  $X \sim N(0, 1)$  and  $Y \sim N(0, 1)$  be independent random variables.

Show that  $X + Y$  is independent of  $X - Y$ .

43. If  $X \sim N(\mu, 9^2)$  and  $Y \sim N(\mu, 12^2)$  are independent, and if

$$P(X + 2Y \leq 3) = P(2X - Y \geq 4), \text{ determine } \mu.$$

[Calcutta Univ. B.Sc. (Maths Hons.), 1989]

44. If  $X \sim N(0, 1)$  and  $Y \sim N(0, 1)$ , prove that

$$(i) \text{Var}(\sin X) > \text{Var}(\cos X)$$

$$(ii) E|X - Y| \leq \sqrt{8/\pi}$$

[Delhi Univ. B.A. Hons. (Spl. Course-Statistics), 1988]

**Hint.** (i)  $X \sim N(0, 1) \Rightarrow \varphi_X(t) = E[\cos tX + i \sin tX] = e^{-t^2/2}$ .

$$\Rightarrow E(\cos tX) = e^{-t^2/2} \text{ and } E(\sin tX) = 0.$$

Taking  $t = 1$  and  $2$ , we get:

$$E(\cos X) = e^{-1/2}; E(\cos 2X) = e^{-2}; E(\sin X) = E(\sin 2X) = 0.$$

$$\begin{aligned} \text{Var}(\cos X) &= E(\cos^2 X) - (E \cos X)^2 = E\left[\frac{1 + \cos 2X}{2}\right] - [E \cos X]^2 \\ &= \frac{1}{2}(1 - e^{-1})^2 \approx 0.99 \end{aligned}$$

$$\text{Similarly } \text{Var}(\sin X) = E\left[\frac{1 - \cos 2X}{2}\right] - [E \sin X]^2 = \frac{1}{2}(1 - e^{-2}) \approx 0.43$$

$$(ii) \text{ Use } |X - Y| \leq |X| + |Y| \text{ and } E|X| = E|Y| = \sqrt{2/\pi}$$

$$\text{or } X - Y \sim N(0, \sigma^2 = 2); \quad E|X - Y| = \sqrt{2/\pi} \sigma = \sqrt{4/\pi} < \sqrt{(8/\pi)}$$

45. Let  $X$  and  $Y$  be independent  $N(0, 1)$  variates. Let  $X = R \cos \theta$ ,  $Y = R \sin \theta$ . Show that  $R$  and  $\theta$  are independent variates.

[Delhi Univ. B.A. Hons. (Spl. Course Statistics), 1985]

46. If  $X \sim N(0, 1)$ , find p.d.f. of  $|X|$ . Hence or otherwise evaluate  $E|X|$ . [Delhi Univ. B.Sc. (Maths. Hons.), 1980]

**Hint.** Distribution function  $G_Y(y)$  of  $Y = |X|$  is given by:

$$\begin{aligned} G_Y(y) &= P(Y \leq y) = P(|X| \leq y) = P(-y \leq X \leq y) \\ &= P(X \leq y) - P(X \leq -y) \end{aligned}$$

$$G_Y(y) = F_X(y) - F_X(-y).$$

where  $F(\cdot)$  is the distribution function of  $X$ . Differentiating, the p.d.f. of  $Y = |X|$  is given by

$$g_Y(y) = f_X(y) + f_X(-y) = 2f_X(y)$$

$$\Rightarrow g_Y(y) = \sqrt{2/\pi} \cdot e^{-y^2/2}; y \geq 0 \quad [\text{By symmetry, since } X \sim N(0, 1)]$$

**8.2.15. The log-normal Distribution.** The positive r.v.  $X$  is said to have a log-normal distribution if  $\log_e X$  is normally distributed.

Let  $Y = \log_e X \sim N(\mu, \sigma^2)$ .

For  $x > 0$ ,

$$F_X(x) = P(X \leq x) = P(\log X \leq \log x) = P(Y \leq \log x) \quad (\text{since } \log X \text{ is monotonic increasing function})$$

$$\begin{aligned} &= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\log x} \exp\left\{-\left(y - \mu\right)^2 / 2 \sigma^2\right\} dy \\ &= \frac{1}{\sigma \sqrt{2\pi}} \int_0^x \exp\left\{-\left(\log u - \mu\right)^2 / 2 \sigma^2\right\} \frac{du}{u} \end{aligned} \quad [ \text{since } Y \sim N(\mu, \sigma^2) ] \quad (y = \log u)$$

For  $x \leq 0$ ,

$$F_X(x) = P(X \leq x) = 0$$

Let us define

$$f_X(u) = \begin{cases} \frac{1}{u \sigma \sqrt{2\pi}} \cdot \exp\left\{-\left(\log u - \mu\right)^2 / 2 \sigma^2\right\}, & u > 0 \\ 0, & u \leq 0 \end{cases} \quad \dots(8.18)$$

Then  $F_X(x) = \int_{-\infty}^x f_X(u) du$  for every  $x$  and hence  $f_X(x)$  defined in (8.18)

is a p.d.f. of  $X$ .

**Remark.** If  $X \sim N(\mu, \sigma^2)$ , then  $Y = e^X$  is called a log-normal random variable, since its logarithm  $\log Y = X$ , is a normal r.v.

**Moments.** The  $r$ th moment about origin is given by

$$\begin{aligned} \mu'_r &= E(X^r) = E(e^{rY}) && [\because Y = \log X \Rightarrow X = e^Y] \\ &= M_Y(r) && (\text{m.g.f. of } Y, r \text{ being the parameter}) \\ &= \exp\left\{\mu r + \frac{1}{2} r^2 \sigma^2\right\} && [\because Y \sim N(\mu, \sigma^2)] \end{aligned} \quad \dots(8.19)$$

**Remarks. 1.** In particular if we take  $\mu = \log \alpha$ ,  $\alpha > 0$  i.e.,  
 $\log X \sim N(\log \alpha, \sigma^2)$ , then

$$\mu'_r = E(X^r) = \exp\left\{r \cdot \log \alpha + \frac{1}{2} r^2 \sigma^2\right\} = \alpha^r \cdot \exp\left\{r^2 \sigma^2 / 2\right\} \quad \dots(8.19a)$$

$$\therefore \text{mean} = \mu'_1 = \alpha e^{\sigma^2 / 2} \text{ and } \mu'_2 = \alpha^2 e^{\sigma^2}$$

$$\mu'_2 = \mu'_2 - \mu'_1^2 = \alpha^2 e^{\sigma^2} (e^{\sigma^2} - 1)$$

**2.** It arises in problems of economics, biology, geology, and reliability theory. In particular it arises in the study of dimensions of particles under pulverisation.

**3.** If  $X_1, X_2, \dots, X_n$  is a set of independently identically distributed random variables such that mean of each  $\log X_i$  is  $\mu$  and its variance is  $\sigma^2$ , then the product  $X_1 X_2 \dots X_n$  is asymptotically distributed according to logarithmic normal distribution and with mean  $\mu$  and variance  $n \sigma^2$

## EXERCISE 8(c)

1. (a) Let  $X$  be a non-negative random variable such that  $\log X = Y$ , (say), is normally distributed with mean  $\mu$  and variance  $\sigma^2$ .

(i) Write down the probability density function of  $X$ . Find  $E(X)$  and  $\text{Var}(X)$ .

(ii) Find the median and the mode of the distribution of  $X$ .

(b) If  $X$  is a normally distributed with zero mean and variance  $\sigma^2$  find the density function of  $U = e^X$ . Locate the mode of the distribution.

2. A random variable  $X$  has the probability density function:

$$f(x) = \frac{1}{\beta x \sqrt{2\pi}} \exp\left[-\frac{1}{2\beta^2}(\log x - \alpha)^2\right], \quad x > 0.$$

Find  $E(X)$  and  $\text{Var}(X)$ .

[Punjab Univ. M.A. (Eco.), 1991].

3. A random variate  $X$  has the p.d.f.,

$$f(x) = \begin{cases} \frac{1}{x \sqrt{2\pi}} e^{-(\log x)^2/2}, & x > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Calculate the mean, mode, standard deviation and coefficient of skewness.

Ans.  $\sqrt{e}$ ,  $1/e$ ,  $\sqrt{e(e-1)}$ , and  $(1-e^{-3/2})/\sqrt{e-1}$

4. The random variable  $X$  has mean  $m$  and standard deviation  $s$ . If  $\gamma = \log X$  is normally distributed with mean  $M$  and standard deviation  $S$ , prove that

$$(i) \quad m = \exp\left[M + \frac{1}{2}S^2\right], \quad (ii) \quad 1 + \frac{s^2}{m^2} = e^{S^2}$$

5. Given that  $X_i$  are independent logarithmic normal variates with parameters  $\mu_i$  and  $\sigma_i$ ;  $i = 1, 2, \dots, n$ , find the  $s$ th raw moment of the variable

$$Y = \prod (a_i X_i); \quad i = 1, 2, \dots, n$$

6. Show that the log-normal distribution is positively skewed i.e., mean > median > mode.

Ans. Let  $Y = \log X \sim N(\mu, \sigma^2)$

$$E(X) = e^{\mu + \sigma^2/2}; \quad \text{Median} = e^\mu; \quad \text{Mode} = e^{\mu - \sigma^2}$$

7. If  $X$  and  $Y$  are two independent log-normal variates, then  $XY$  and  $X/Y$  are also log-normal variates.

Hint. Let  $\log X \sim N(\mu_1, \sigma_1^2)$ ;  $\log Y \sim N(\mu_2, \sigma_2^2)$ ;  $U = XY$  and  $V = (X/Y)$ .

$$\begin{aligned} \log U &= \log X + \log Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\ \log V &= \log X - \log Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) \end{aligned} \quad \left\{ \because X \text{ and } Y \text{ are independent} \right.$$

8. If  $X \sim N(0, \sigma^2)$ , obtain the distribution of  $e^X$ . Find out the mean of the distribution.

[Delhi Univ. B.Sc. (Stat. Hons.), 1985]

9. If  $X \sim N(\mu, \sigma^2)$ , find the p.d.f. of  $Y = e^X$ , using the result that  $E(e^{tX}) = e^{\mu t + t^2 \sigma^2/2}$ .

Find the coefficient of variation of  $Y$ . [Delhi Univ. M.A. (Eco.), 1991]

**8.3. Gamma Distribution.** The continuous random variable  $X$  which is distributed according to the probability law :

$$f(x) = \begin{cases} \frac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)}; & \lambda > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases} \quad \dots(8.20)$$

is known as a Gamma variate with parameter  $\lambda$  and referred to as a  $\gamma(\lambda)$  variate and its distribution is called the Gamma-distribution.

**Remarks.** 1. The function  $f(x)$  defined above represents a probability function, since

$$\int_0^\infty f(x) dx = \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{-x} x^{\lambda-1} dx = \frac{1}{\Gamma(\lambda)} \cdot \Gamma(\lambda) = 1$$

2. A continuous random variable  $X$  having the following p.d.f. is said to have a gamma distribution with two parameters  $a$  and  $\lambda$ .

$$f(x) = \left. \begin{cases} \frac{a^\lambda}{\Gamma(\lambda)} e^{-ax} x^{\lambda-1}; & a > 0, \lambda > 0; 0 < x < \infty \\ 0, & \text{otherwise} \end{cases} \right\} \quad \dots(8.20a)$$

We write  $X \sim \gamma(a, \lambda)$

Taking  $a = 1$  in (8.20a) we get (8.20). Hence we may write

$$X \sim \gamma(\lambda) = (1, \lambda).$$

3. The cumulative distribution function, called incomplete gamma function is

$$F_X(x) = \begin{cases} \int_0^x f(u) du = \frac{1}{\Gamma(\lambda)} \int_0^x e^{-u} u^{\lambda-1} du, & x > 0 \\ 0, & \text{otherwise} \end{cases} \quad \dots(8.20b)$$

**8.3.1. M.G.F. of Gamma Distribution.** M.G.F. about origin is given by

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \int_0^\infty e^{tx} f(x) dx = \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{tx} e^{-x} x^{\lambda-1} dx \\ &= \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{-(1-t)x} x^{\lambda-1} dx = \frac{1}{\Gamma(\lambda)} \cdot \frac{\Gamma(\lambda)}{(1-t)^\lambda}, |t| < 1 \end{aligned}$$

$$\therefore M_X(t) = (1-t)^{-\lambda}, |t| < 1 \quad \dots(8.21)$$

**8.3.2. Cumulant Generating Function of Gamma Distribution.** The cumulant generating function  $K_X(t)$  is given by

$$K_X(t) = \log M_X(t) = \log (1-t)^{-\lambda} = -\lambda \log (1-t); |t| < 1$$

$$= \lambda \left[ t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \dots \right]$$

$\therefore$  Mean =  $\kappa_1$  = Coefficient of  $t$  in  $K_X(t) = \lambda$

$$\mu_2 = \kappa_2 = \text{Coefficient of } \frac{t^2}{2!} \text{ in } K_X(t) = \lambda$$

$$\kappa_3 = \text{Coefficient of } \frac{t^3}{3!} \text{ in } K_X(t) = 2\lambda$$

$$\kappa_4 = \text{Coefficient of } \frac{t^4}{4!} \text{ in } K_X(t) = 6\lambda$$

$$\therefore \mu_4 = \kappa_4 + 3\kappa_2^2 = 6\lambda + 3\lambda^2$$

$$\text{Hence } \beta_1 = \frac{\mu_3}{\mu_2^2} = \frac{4\lambda^2}{\lambda^3} = \frac{4}{\lambda}, \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{6}{\lambda}$$

**Remarks 1.** Like Poisson distribution, the mean and variance of the Gamma distribution are also equal. However, Poisson distribution is discrete while Gamma distribution is continuous.

**2. Limiting form of Gamma distribution as  $\lambda \rightarrow \infty$ .** We know that if  $X \sim \gamma(\lambda)$ , then  $E(X) = \lambda = \mu$ , (say), and  $\text{Var}(X) = \lambda = \sigma^2$ , (say). Then standard gamma variate is given by

$$Z = \frac{X - \mu}{\sigma} = \frac{X - \lambda}{\sqrt{\lambda}}$$

$$M_Z(t) = e^{-\mu t/\sigma} M_X(t/\sigma) = e^{-\mu t/\sigma} (1 - t/\sigma)^{-\lambda}. \quad [\text{From (8.21)}].$$

$$\begin{aligned} &= e^{-t\lambda/\sqrt{\lambda}} \left( 1 - \frac{t}{\sqrt{\lambda}} \right)^{-\lambda} \\ \Rightarrow K_Z(t) &= -\sqrt{\lambda} \cdot t - \lambda \log \left( 1 - \frac{t}{\sqrt{\lambda}} \right) \\ &= -\sqrt{\lambda} \cdot t + \lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} + \frac{t^3}{3\lambda^{3/2}} + \dots \right) \\ &= -\sqrt{\lambda} \cdot t + \sqrt{\lambda} \cdot t + \frac{t^2}{2} + o(\lambda^{-1/2}) \end{aligned}$$

where  $o(\lambda^{-1/2})$  are terms containing  $\frac{1}{2}$  and higher powers of  $\lambda$  in the denominator.

$$\therefore \lim_{\lambda \rightarrow \infty} K_Z(t) = \frac{t^2}{2} \Rightarrow \lim_{\lambda \rightarrow \infty} M_Z(t) = e^{t^2/2},$$

which is the m.g.f. of a Standard Normal Variate. Hence by uniqueness theorem of m.g.f., Standard Gamma variate tends to Standard Normal Variate as  $\lambda \rightarrow \infty$ . In other words, Gamma distribution tends to Normal distribution for large values of parameter  $\lambda$ .

**3. For the two parameter gamma distribution (8.20 a), we have**

$$M_X(t) = \left(1 - \frac{t}{a}\right)^{-\lambda} ; t < a . \quad \dots(8.21a)$$

Proof is left as an exercise to the reader.

$$K_X(t) = -\lambda \log(1-t/a) ; t < a$$

$$= \lambda \left[ \frac{t}{a} + \frac{1}{2} \left( \frac{t}{a} \right)^2 + \frac{1}{3} \left( \frac{t}{a} \right)^3 + \dots \right]$$

$$\therefore \text{Mean} = k_1 = \lambda/a$$

$$\text{Variance} = k_2 = \lambda/a^2 = \text{Mean}/a \quad \dots(8.21b)$$

Hence Variance > Mean if  $a < 1$ ,

Variance = Mean if  $a = 1$ ,

and Variance < Mean if  $a > 1$ .

**8.3.3. Additive Property of Gamma Distribution.** *The sum of independent Gamma variates is also a Gamma variate. More precisely, if  $X_1, X_2, \dots, X_k$  are independent Gamma variates with parameters  $\lambda_1, \lambda_2, \dots, \lambda_k$ , respectively then  $X_1 + X_2 + \dots + X_k$  is also a Gamma variate with parameter  $\lambda_1 + \lambda_2 + \dots + \lambda_k$ .*

**Proof.** Since  $X_i$  is a  $\gamma(\lambda_i)$  variate,

$$M_{X_i}(t) = (1-t)^{-\lambda_i}$$

The m.g.f. of the sum  $X_1 + X_2 + \dots + X_k$  is given by

$$M_{X_1+X_2+\dots+X_k}(t) = M_{X_1}(t) M_{X_2}(t) \dots M_{X_k}(t)$$

(since  $X_1, X_2, \dots, X_k$  are independent)

$$= (1-t)^{-\lambda_1} (1-t)^{-\lambda_2} \dots (1-t)^{-\lambda_k}$$

$$= (1-t)^{-(\lambda_1 + \lambda_2 + \dots + \lambda_k)}$$

which is the m.g.f. of a Gamma variate with parameter  $\lambda_1 + \lambda_2 + \dots + \lambda_k$ . Hence the result follows by the uniqueness theorem of m.g.f.'s.

**Remark.** If general, if  $X_i \sim \gamma(a, \lambda_i)$ ,  $i = 1, 2, \dots, n$  are independent r.v.'s

$$\text{then } \sum_{i=1}^n X_i \sim \gamma\left(a, \sum_{i=1}^n \lambda_i\right).$$

**8.4. Beta Distribution of First Kind.** *The continuous random variable which is distributed according to the probability law*

$$f(x) = \begin{cases} \frac{1}{B(\mu, \nu)} \cdot x^{\mu-1} (1-x)^{\nu-1} ; (\mu, \nu) > 0, 0 < x < 1 \\ 0, \text{ otherwise} \end{cases} \quad \dots(8.22)$$

(where  $B(\mu, \nu)$  is the Beta function), is known as a Beta variate of the first kind with parameters  $\mu$  and  $\nu$  and is referred to as  $\beta_1(\mu, \nu)$  variate and its distribution is called Beta distribution of the first kind.

**Remarks. 1.** The cumulative distribution function, often called the Incomplete Beta Function, is

$$F(x) = \begin{cases} 0, & x < 0 \\ \int_0^x \frac{1}{B(\mu, \nu)} u^{\mu-1} (1-u)^{\nu-1} du, & 0 < x < 1, (\mu, \nu) > 0 \\ 1, & x > 1 \end{cases} \quad \dots(8.22\ a)$$

2. In particular, if we take  $\mu = 1$  and  $\nu = 1$  in (8.22) we get:

$$f(x) = \frac{1}{\beta(1, 1)} = 1, \quad 0 < x < 1 \quad \dots(8.22\ a)$$

which is the p.d.f. of uniform distribution on  $[0, 1]$ .

3. If  $X \sim \beta_1(\mu, \nu)$ , then it can be easily proved that  $1 - X \sim \beta_1(\nu, \mu)$ .

#### 8.4.1. Constants of Beta Distribution of First Kind.

$$\begin{aligned} \mu_r' &= \int_0^1 x^r f(x) dx = \frac{1}{B(\mu, \nu)} \int_0^1 x^{\mu+r-1} (1-x)^{\nu-1} dx \\ &= \frac{1}{B(\mu, \nu)} B(\mu+r, \nu) = \frac{\Gamma(\mu+r)\Gamma(\nu)}{\Gamma(\mu+r+\nu)} \cdot \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)} \\ &= \frac{\Gamma(\mu+r)\Gamma(\mu+\nu)}{\Gamma(\mu+r+\nu)\Gamma(\mu)} \end{aligned} \quad \dots(8.22\ b)$$

In particular

$$\text{Mean } \mu_1' = \frac{\Gamma(\mu+1)}{\Gamma(\mu+\nu+1)} \cdot \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)} = \frac{\mu \Gamma(\mu) \Gamma(\mu+\nu)}{(\mu+\nu) \Gamma(\mu+\nu) \Gamma(\mu)} = \frac{\mu}{\mu+\nu} \quad \dots(8.22\ c)$$

[  $\because \Gamma(k) = (k-1) \Gamma(k-1)$  ]

$$\begin{aligned} \mu_2' &= \frac{\Gamma(\mu+2) \cdot \Gamma(\mu+\nu)}{\Gamma(\mu+\nu+2) \Gamma(\mu)} = \frac{(\mu+1) \mu \Gamma(\mu) \Gamma(\mu+\nu)}{(\mu+\nu+1) (\mu+\nu) \Gamma(\mu+\nu) \Gamma(\mu)} \\ &= \frac{\mu(1+\mu)}{(\mu+\nu)(\mu+\nu+1)} \end{aligned}$$

$$\begin{aligned} \text{Hence } \mu_2 &= \mu_2' - \mu_1'^2 = \frac{\mu(1+\mu)}{(\mu+\nu)(\mu+\nu+1)} - \left( \frac{\mu}{\mu+\nu} \right)^2 \\ &= \frac{\mu}{(\mu+\nu)^2(\mu+\nu+1)} \left[ (\mu+\nu)(\mu+1) - \mu(\mu+\nu+1) \right] \\ &= \frac{\mu\nu}{(\mu+\nu)^2(\mu+\nu+1)} \end{aligned} \quad \dots(8.22\ d)$$

Similarly, we have

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \frac{2\mu\nu(\nu-\mu)}{(\mu+\nu)^3(\mu+\nu+1)(\mu+\nu+2)}$$

and

$$\begin{aligned} \mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 \\ &= \frac{3\mu\nu\{\mu\nu(\mu+\nu-6) + 2(\mu+\nu)^2\}}{(\mu+\nu)^4(\mu+\nu+1)(\mu+\nu+2)(\mu+\nu+3)} \end{aligned}$$

so that

$$\beta_1 = \frac{\mu^2}{\mu^2} = \frac{4(v-\mu)^2(\mu+v+1)}{\mu v (\mu+v+2)^2}$$

and  $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3(\mu+v+1) \cdot \mu v (\mu+v-6) + 2(\mu+v)^2}{\mu v (\mu+v+2)(\mu+v+3)}$

The harmonic mean  $H$  is given by

$$\begin{aligned} \frac{1}{H} &= \int_0^1 \frac{1}{x} f(x) dx = \frac{1}{B(\mu, v)} \int_0^1 x^{\mu-2} (1-x)^{v-1} dx \\ &= \frac{1}{B(\mu, v)} B(\mu-1, v) = \frac{\Gamma(\mu-1) \Gamma(v)}{\Gamma(\mu+v-1)} \cdot \frac{\Gamma(\mu+v)}{\Gamma(\mu) \Gamma(v)} \\ &= \frac{\Gamma(\mu-1)(\mu+v-1) \Gamma(\mu+v-1)}{\Gamma(\mu+v-1)(\mu-1) \Gamma(\mu-1)} = \frac{\mu+v-1}{\mu-1} \\ \therefore H &= \frac{\mu-1}{\mu+v-1} \end{aligned} \quad \dots(8.22e)$$

**8.5. Beta Distribution of Second Kind.** The continuous random variable  $X$  which is distributed according to the probability law :

$$f(x) = \begin{cases} \frac{1}{B(\mu, v)} \cdot \frac{x^{\mu-1}}{(1+x)^{\mu+v}} ; & (\mu, v) > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases} \quad \dots(8.23)$$

is known as a Beta variate of second kind with parameters  $\mu$  and  $v$  and is denoted as  $\beta_2(\mu, v)$  variate and its distribution is called Beta distributed of second kind.

**Remark.** Beta distribution of second kind is transformed to Beta distribution of first kind by the transformation

$$1+x = \frac{1}{y} \Rightarrow y = \frac{1}{1+x} \quad \dots(*)$$

Thus if  $X \sim \beta_2(\mu, v)$ , then  $Y$  defined in (\*) is a  $\beta_1(\mu, v)$ . The proof is left as an exercise to the reader.

#### 8.5.1. Constants of Beta Distribution of Second Kind.

$$\begin{aligned} \mu_r' &= \int_0^\infty x^r f(x) dx = \frac{1}{B(\mu, v)} \int_0^\infty \frac{x^{\mu+r-1}}{(1+x)^{\mu+v}} dx \\ &= \frac{1}{B(\mu, v)} \int_0^\infty \frac{x^{(\mu+r)-1}}{(1+x)^{\mu+r+v-r}} dx \\ &= \frac{1}{B(\mu, v)} \cdot B(\mu+r, v-r) \\ &= \frac{\Gamma(\mu+r) \Gamma(v-r)}{\Gamma(\mu+v)} \cdot \frac{\Gamma(\mu+v)}{\Gamma(\mu) \Gamma(v)} = \frac{\Gamma(\mu+r) \Gamma(v-r)}{\Gamma(\mu) \Gamma(v)} \end{aligned} \quad \dots(1)$$

In particular

$$\mu_1' = \frac{\Gamma(\mu+1)\Gamma(v-1)}{\Gamma(\mu)\Gamma(v)} = \frac{\mu\Gamma(\mu)\Gamma(v-1)}{\Gamma(\mu)(v-1)\Gamma(v-1)} = \frac{\mu}{v-1}$$

$$\mu_2' = \frac{\Gamma(\mu+2)\Gamma(v-2)}{\Gamma(\mu)\Gamma(v)} = \frac{(\mu+1)\mu\Gamma(\mu)\Gamma(v-2)}{\Gamma(\mu)(v-1)(v-2)\Gamma(v-2)} = \frac{\mu(\mu+1)}{(v-1)(v-2)}$$

$$\begin{aligned}\mu_2 &= \mu_2' - \mu_1'^2 = \frac{\mu(\mu+1)}{(v-1)(v-2)} - \left( \frac{\mu}{v-1} \right)^2 \\ &= \frac{\mu}{v-1} \left[ \frac{(v-1)(\mu+1) - \mu(v-2)}{(v-1)(v-2)} \right] = \frac{\mu(\mu+v-1)}{(v-1)^2(v-2)}\end{aligned}$$

The harmonic mean  $H$  is given by

$$\begin{aligned}\frac{1}{H} &= E\left(\frac{1}{X}\right) = \int_0^\infty \frac{1}{x} \cdot f(x) dx = \frac{1}{B(\mu, v)} \int_0^\infty \frac{x^{\mu-2}}{(1+x)^{\mu+v}} dx \\ &= \frac{1}{B(\mu, v)} \int_0^\infty \frac{x^{\mu-1-1}}{(1+x)^{\mu-1+v+1}} dx = \frac{1}{B(\mu, v)} \cdot B(\mu-1, v+1), \mu > 1. \\ &= \frac{\Gamma(\mu-1)\Gamma(v+1)}{\Gamma(\mu+v)} \cdot \frac{\Gamma(\mu+v)}{\Gamma(v)\Gamma(v)} = \frac{\Gamma(\mu-1)v\Gamma(v)}{(\mu-1)\Gamma(\mu-1)\Gamma(v)} = \frac{v}{\mu-1}\end{aligned}$$

$$\text{Hence } H = \frac{\mu-1}{v}$$

**Example 8.29.** The daily consumption of milk in a city, in excess of 20,000 gallons, is approximately distributed as a Gamma variate with the parameters  $v = 2$  and  $\lambda = \frac{1}{10,000}$ . The city has a daily stock of 30,000 gallons. What is the probability that the stock is insufficient on a particular day?

[Madras Univ. B.Sc. (Stat. Main), 1990]

**Solution.** If the r.v.  $X$  denotes the daily consumption of milk (in gallons) in a city, then the r.v.  $Y = X - 20,000$  has a gamma distribution with p.d.f.

$$g(y) = \frac{1}{(10,000)^2 \Gamma(2)} y^{2-1} e^{-y/10,000} = \frac{y e^{-y/10,000}}{(10,000)^2}; 0 < y < \infty$$

[See 8.20 (a)]

Since the daily stock of the city is 30,000 gallons, the required probability ' $p$ ' that the stock is insufficient on a particular day is given by

$$p = P(X > 30,000) = P(Y > 10,000)$$

$$\begin{aligned}&= \int_{10,000}^{\infty} g(y) dy = \int_{10,000}^{\infty} \frac{y e^{-y/10,000}}{(10,000)^2} dy \\ &= \int_1^{\infty} z e^{-z} dz\end{aligned}$$

[Taking  $z = y/10,000$ ]

Integrating by parts, we get

$$P = \left| -z e^{-z} \right|_1^\infty + \int_1^\infty e^{-z} dz = e^{-1} - \left| e^{-z} \right|_1^\infty \\ = e^{-1} + e^{-1} = 2/e$$

**Remark.** Since  $v = 2$ , the integration is easily done. However, for general values of  $\lambda$  and  $v$ , the integral is evaluated by using tables of Incomplete Gamma Integral, [see Tables of Incomplete Gamma Functions, K. Pearson; Cambridge University Press] of the form

$$\int_0^{\alpha} \frac{e^{-x} \cdot x^{n-1}}{\Gamma n} \cdot dx,$$

which have been tabulated for different values of  $\alpha$  and  $n$ .

**Example 8.30.** If  $X \sim N(\mu, \sigma^2)$ , obtain the p.d.f. of:

$$U = \frac{1}{2} \left( \frac{X-\mu}{\sigma} \right)^2$$

**Solution.** Since  $X \sim N(\mu, \sigma^2)$ ,

$$dP(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx, -\infty < x < \infty$$

$$\text{Let } u = \frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \Rightarrow \frac{x-\mu}{\sigma} = \sqrt{2u}$$

$$\therefore dx = \frac{\sigma}{\sqrt{2u}} du$$

Hence probability differential of  $U$  is

$$dG(u) = \frac{1}{\sqrt{2\pi}\sigma} e^{-u} \cdot \frac{\sigma}{\sqrt{2u}} du = \frac{1}{2\sqrt{\pi}} e^{-u} u^{-1/2} du$$

$$= \frac{1}{\sqrt{\pi}} e^{-u} u^{-1/2} du, 0 < u < \infty$$

the factor  $\frac{1}{2}$  disappearing from the fact that total probability is unity.

$$\therefore dG(u) = \frac{1}{\Gamma(\frac{1}{2})} e^{-u} u^{(1/2)-1} du, 0 < u < \infty \quad [\because \Gamma(\frac{1}{2}) = \sqrt{\pi}]$$

Hence  $U = \frac{1}{2} \left( \frac{X-\mu}{\sigma} \right)^2$  is a  $\gamma(\frac{1}{2})$  variate.

**Example 8.31.** Show that the mean value of positive square root of a  $\gamma(\mu)$  variate is  $\Gamma(\mu + \frac{1}{2})/\Gamma(\mu)$ . Hence prove that the mean deviation of a normal variate from its mean is  $\sqrt{2/\pi} \sigma$ , where  $\sigma$  is the standard deviation of the distribution.

[Delhi Univ. B.Sc. (Stat. Hons.), 1986]

**Solution.** Let  $X$  be a  $\gamma(\mu)$  variate. Then

$$f(x) = \frac{e^{-x} x^{\mu-1}}{\Gamma(\mu)} ; \mu > 0, 0 < x < \infty$$

$$\therefore E(\sqrt{X}) = \int_0^{\infty} x^{1/2} f(x) dx = \frac{1}{\Gamma(\mu)} \int_0^{\infty} e^{-x} x^{\mu + (1/2) - 1} dx = \frac{\Gamma(\mu + \frac{1}{2})}{\Gamma(\mu)}$$

If  $X \sim N(\mu, \sigma^2)$ , then

$$U = \frac{1}{2} \left( \frac{X - \mu}{\sigma} \right)^2 \text{ is a } \gamma\left(\frac{1}{2}\right) \text{ variate.} \quad (\text{c.f. Example 8-30})$$

$$\therefore |X - \mu| = \sqrt{2} \sigma U^{1/2}, \text{ where } U \text{ is a } \gamma\left(\frac{1}{2}\right) \text{ variate.}$$

Hence mean deviation about mean is given by

$$\begin{aligned} E|X - \mu| &= E(\sqrt{2} \sigma U^{1/2}) = \sqrt{2} \sigma E(U^{1/2}) \\ &= \sqrt{2} \sigma \frac{\Gamma\left(\frac{1}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} = \frac{\sqrt{2} \sigma}{\sqrt{\pi}} = \sqrt{2/\pi} \sigma \end{aligned}$$

**Example 8-32.** If  $X$  and  $Y$  are independent Gamma variates with parameters  $\mu$  and  $v$  respectively, show that the variables

$$U = X + Y, Z = \frac{X}{X + Y}$$

are independent and that  $U$  is a  $\gamma(\mu + v)$  variate and  $Z$  is a  $\beta_1(\mu, v)$  variate.

[Delhi Univ. B.Sc. (Stat. Hons.), 1991]

**Solution.** Since  $X$  is a  $\gamma(\mu)$  variate and  $Y$  is a  $\gamma(v)$  variate, we have

$$f_1(x) dx = \frac{1}{\Gamma(\mu)} e^{-x} x^{\mu-1} dx; 0 < x < \infty, \mu > 0$$

$$f_2(y) dy = \frac{1}{\Gamma(v)} e^{-y} y^{v-1} dy; 0 < y < \infty, v > 0$$

Since  $X$  and  $Y$  are independently distributed, their joint probability differential is given by the compound probability theorem as

$$dF(x, y) = f_1(x) f_2(y) dx dy = \frac{1}{\Gamma(\mu) \Gamma(v)} e^{-(x+y)} x^{\mu-1} y^{v-1} dx dy$$

$$\text{Now } u = x + y, z = \frac{x}{x + y}, \text{ so that } x = uz, y = u - x = u(1 - z)$$

Jacobian of transformation  $J$  is given by

$$J = \frac{\partial(x, y)}{\partial(u, z)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial z} & \frac{\partial y}{\partial z} \end{vmatrix} = \begin{vmatrix} z & 1-z \\ u & -u \end{vmatrix} = -u$$

As  $X$  and  $Y$  range from 0 to  $\infty$ ,  $u$  ranges from 0 to  $\infty$  and  $z$  from 0 to 1.

Hence the joint distribution of  $U$  and  $Z$  is

$$\begin{aligned} dG(u, z) &= g(u, z) du dz = \frac{1}{\Gamma(\mu) \Gamma(v)} e^{-u} (uz)^{\mu-1} [u(1-z)]^{v-1} |J| du dz \\ &= \frac{1}{\Gamma(\mu) \Gamma(v)} \cdot e^{-u} u^{\mu+v-1} z^{\mu-1} (1-z)^{v-1} du dz \end{aligned}$$

$$\begin{aligned}
 &= \left( \frac{e^{-u} \cdot u^{\mu+v-1}}{\Gamma(\mu+v)} du \right) \left( \frac{1}{B(\mu, v)} z^{\mu-1} (1-z)^{v-1} dz \right) \\
 &= [g_1(u) du] [g_2(z) dz], \quad (\text{say}),
 \end{aligned}$$

where  $g_1(u) = \frac{1}{\Gamma(\mu+v)} e^{-u} u^{\mu+v-1}, 0 < u < \infty$

and  $g_2(z) = \frac{1}{B(\mu, v)} z^{\mu-1} (1-z)^{v-1}, 0 < z < 1$

From (\*) and (\*\*) we conclude that  $U$  and  $Z$  are independently distributed  $U$  as a  $\gamma(\mu+v)$  variate and  $Z$  as a  $\beta_1(\mu, v)$  variate.

**Example 8.33.** If  $X$  and  $Y$  are independent Gamma variates with parameters  $\mu$  and  $v$  respectively, show that

$$U = X + Y, Z = \frac{X}{Y}$$

are independent and that  $U$  is a  $\gamma(\mu+v)$  variate and  $Z$  is a  $\beta_2(\mu, v)$  variate  
[Rajasthan Univ. B.Sc. (Hons.), 1

**Solution.** As in example 8.32, we have

$$dF(x, y) = \frac{1}{\Gamma(\mu) \Gamma(v)} e^{-(x+y)} x^{\mu-1} y^{v-1} dx dy, 0 < (x, y) < \infty$$

Since  $u = x + y$  and  $z = \frac{x}{y}$ ,

$$1+z = 1 + \frac{x}{y} = \frac{u}{y} \Rightarrow y = \frac{u}{1+z} \text{ and } x = \frac{uz}{1+z} = u \left( 1 - \frac{1}{1+z} \right)$$

$$J = \frac{\partial(x, y)}{\partial(u, z)} = \frac{-u}{(1+z)^2}$$

As  $x$  and  $y$  range from 0 to  $\infty$ , both  $u$  and  $z$  range from 0 to  $\infty$ . Hence the joint probability differential of random variables  $U$  and  $Z$  becomes

$$\begin{aligned}
 dG(u, v) &= \frac{1}{\Gamma(\mu) \Gamma(v)} e^{-u} \left( \frac{uz}{1+z} \right)^{\mu-1} \left( \frac{u}{1+z} \right)^{v-1} |J| du dz \\
 &= \left( \frac{e^{-u} u^{\mu+v-1}}{\Gamma(\mu+v)} du \right) \left( \frac{1}{B(\mu, v)} \cdot \frac{z^{\mu-1}}{(1+z)^{\mu+v}} dv \right); \\
 &\quad 0 < u < \infty, 0 < v < \infty
 \end{aligned}$$

Hence  $U$  and  $Z$  are independently distributed,  $U$  as a  $\gamma(\mu+v)$  variate and  $Z$  as a  $\beta_2(\mu, v)$  variate.

**Remark.** The above two examples lead to the following important result. If  $X$  is a  $\gamma(\mu)$  variate and  $Y$  is an independent  $\gamma(v)$  variate, then

(i)  $X + Y$  is a  $\gamma(\mu+v)$  variate, i.e., the sum of two independent Gamma variates is also a Gamma variate.

(ii)  $\frac{X}{Y}$  is a  $\beta_2(\mu, \nu)$ - variate, i.e., the ratio of two independent Gamma variates is also a gamma variate.

(iii)  $\frac{X}{X+Y}$  is a  $\beta_1(\mu, \nu)$  variate.

**Example 8-34.** Let  $X$  and  $Y$  have joint p.d.f.

$$g(x, y) = \frac{e^{-(x+y)} x^3 y^4}{\Gamma 4 \Gamma 5}, x > 0, y > 0 \\ = 0, \text{ elsewhere.}$$

Find (i) p.d.f. of  $U = \frac{X}{X+Y}$ , (ii)  $E(U)$  and (iii)  $E[U - E(U)]^2$

(Allahabad Univ. B.Sc. 1992)

**Solution.** Let  $u = \frac{x}{x+y}$  and  $v = x+y$

$$\Rightarrow x = uv, y = v - x = v - uv = v(1-u)$$

Jacobian of transformation is

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

Hence joint p.d.f. of  $U$  and  $V$  becomes :

$$p(u, v) = g(x, y) \cdot |J| \\ = \frac{1}{\Gamma 4 \Gamma 5} e^{-v} \cdot (uv)^3 [v(1-u)]^4 \times v \\ = \frac{1}{\Gamma 4 \Gamma 5} e^{-v} \cdot v^8 \cdot u^3 (1-u)^4 ; 0 \leq u \leq 1 ; v > 0 \\ \left[ \because u = \frac{x}{x+y} < 1 \text{ and since } x > 0, y > 0 \right. \\ \left. \text{we have } 0 < u < 1 \text{ and } v = x+y \geq 0 \right] \\ = \left[ \frac{1}{\Gamma 9} e^{-v} \cdot v^8 \right] \left[ \frac{\Gamma 9}{\Gamma 4 \Gamma 5} u^3 (1-u)^4 \right]; \\ 0 < u < 1, v > 0$$

$$\Rightarrow p(u, v) = p_1(v) \cdot p_2(u), \quad \dots(*)$$

$$\text{where } p_1(v) = \frac{1}{\Gamma 9} e^{-v} v^8; v > 0$$

$$p_2(u) = \frac{\Gamma 9}{\Gamma 4 \Gamma 5} u^3 (1-u)^4; 0 < u < 1. \quad \dots(**)$$

From (\*), we conclude that  $U$  and  $V$  are independently distributed and from (\*\*), we conclude that

$$U = \frac{X}{X+Y} \sim \beta_1(4, 5).$$

i.e.,  $U$  is a Beta variate of first kind with parameters (4, 5).

**Aliter.** We have

$$\begin{aligned} g(x, y) &= \frac{1}{\Gamma(4)\Gamma(5)} e^{-(x+y)} x^3 y^4 \\ &= \left[ \frac{1}{\Gamma(4)} e^{-x} x^3 \right] \left[ \frac{1}{\Gamma(5)} e^{-y} y^4 \right] \\ &= g_1(x) g_2(y); x > 0, y > 0 \end{aligned}$$

$\Rightarrow X$  and  $Y$  are independently distributed and  $X \sim \gamma(4)$  and  $Y \sim \gamma(5)$

Hence

$$U = \frac{X}{X+Y} \sim \beta_1(4, 5)$$

$$\begin{aligned} \text{Now } E(U) &= \int_0^1 u \cdot p_2(u) du = \frac{1}{B(4, 5)} \cdot \int_0^1 u^4 (1-u)^4 du \\ &= \frac{1}{B(4, 5)} \cdot B(5, 5) \quad [\text{Using Beta integral}] \\ &= \frac{\Gamma 9}{\Gamma 4 \Gamma 5} \times \frac{\Gamma 5 \Gamma 5}{\Gamma 10} = \frac{\Gamma 9 \cdot 4 \Gamma 4}{\Gamma 4 \cdot 9 \Gamma 9} = \frac{4}{9} \end{aligned}$$

$$\begin{aligned} E(U^2) &= \frac{1}{B(4, 5)} \int_0^1 u^2 \cdot u^3 (1-u)^4 du \\ &= \frac{1}{B(4, 5)} \times B(6, 5) = \frac{\Gamma 9}{\Gamma 4 \Gamma 5} \times \frac{\Gamma 6 \Gamma 5}{\Gamma 11} \\ &= \frac{5 \times 4}{10 \times 9} = \frac{2}{9} \end{aligned}$$

$$E[U - E(U)]^2 = E(U^2) - [E(U)]^2 = \frac{2}{9} - \frac{16}{81} = \frac{2}{81}$$

**Example 8.35.** A random sample of size  $n$  is taken from a population with distribution:

$$dP(x) = \frac{1}{\Gamma(\lambda)} e^{-x/a} \left( \frac{x}{a} \right)^{\lambda-1} \frac{dx}{a}; 0 < x < \infty, a > 0, \lambda > 0$$

Find the distribution of the mean  $\bar{X}$ .

**Solution.**

[Delhi Univ. M.Sc. (OR), 1990]

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \int_0^\infty e^{tx} f(x) dx \\ &= \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{tx} e^{-x/a} \left( \frac{x}{a} \right)^{\lambda-1} \frac{dx}{a} \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{\Gamma(\lambda) a^\lambda} \int_0^\infty e^{-\left(\frac{1}{a}-t\right)x} x^{\lambda-1} dx \\
 &= \frac{1}{\Gamma(\lambda) a^\lambda} \cdot \frac{\Gamma(\lambda)}{\left(\frac{1}{a}-t\right)^\lambda} = (1-at)^{-\lambda} \quad \dots(*)
 \end{aligned}$$

$$M_{\bar{X}}(t) = M_{(X_1 + X_2 + \dots + X_n)/n}(t) = M_{X_1 + X_2 + \dots + X_n}(t/n)$$

$$\begin{aligned}
 &[ \because M_{cX}(t) = M_X(ct) ] \\
 &= M_{X_1}(t/n) M_{X_2}(t/n) \dots M_{X_n}(t/n), \\
 &\quad (\text{since } X_1, X_2, \dots, X_n \text{ are independent}).
 \end{aligned}$$

Hence on using (\*), we get

$$M_{\bar{X}}(t) = \left[ \left( 1 - \frac{at}{n} \right)^{-\lambda} \right]^n = \left( 1 - \frac{t}{n/a} \right)^{-n\lambda}$$

which is the m.g.f. of a Gamma distribution (c.f. Remark 3, § 8-3-2). Hence by uniqueness theorem of mg.f.,  $\bar{X} \sim \gamma(n/a, n\lambda)$  with p.d.f.

$$g(\bar{x}) = \frac{(n/a)^{\lambda n}}{\Gamma(\lambda n)} e^{-n\bar{x}/a} (\bar{x})^{n\lambda-1}, \quad 0 < \bar{x} < \infty$$

**Example 8-36.** A sample of  $n$  values is drawn from a population whose probability density is  $ae^{-ax}$ , ( $x \geq 0, a > 0$ ). If  $\bar{X}$  is mean of the sample, show that  $na\bar{X}$  is a  $\gamma(n)$  variate and prove that

$$E(\bar{X}) = \frac{1}{a} \text{ and S.E. of } \bar{X} = \frac{1}{a\sqrt{n}}$$

(Marathwada Univ. M.A., 1991)

**Solution.**

$$f(x) = ae^{-ax}; \quad 0 \leq x \leq \infty, a > 0$$

$$\begin{aligned}
 \therefore M_X(t) &= \int_0^\infty e^{tx} f(x) dx = a \int_0^\infty e^{-(a-t)x} dx \\
 &= a \left| \frac{e^{-(a-t)x}}{-(a-t)} \right|_0^\infty = \frac{a}{a-t}, \quad (a > t)
 \end{aligned}$$

$$\begin{aligned}
 \therefore an\bar{X} &= an \left( \frac{X_1 + X_2 + \dots + X_n}{n} \right) = a(X_1 + X_2 + \dots + X_n)
 \end{aligned}$$

$$\begin{aligned}
 \therefore M_{a\bar{X}}(t) &= M_a(X_1 + X_2 + \dots + X_n)(t) = M_{X_1 + X_2 + \dots + X_n}(at) \\
 &= M_{X_1}(at) \cdot M_{X_2}(at) \dots M_{X_n}(at), \\
 &\quad (\text{since the sample values are independent}).
 \end{aligned}$$

$$\therefore M_{an\bar{X}}(t) = \prod_{i=1}^n M_{X_i}(at) = [M_{X_i}(at)]^n$$

(since  $X_1, X_2, \dots, X_n$  are identically distributed).

$$\therefore M_{an\bar{X}}(t) = \left( \frac{1}{1-t} \right)^n = (1-t)^{-n}, \text{ which is the mg.f. of a } \gamma(n) \text{ variate.}$$

Hence by uniqueness theorem of m.g.f.,  $an\bar{X}$  is a  $\gamma(n)$  variate.

Since the mean and variance of a  $\gamma(n)$  variate are equal, each being equal to  $n$ , we have

$$E(an\bar{X}) = n \Rightarrow an E(\bar{X}) = n, \text{ i.e., } E(\bar{X}) = \frac{1}{a}$$

$$\text{and } V(an\bar{X}) = n \Rightarrow a^2 n^2 V(\bar{X}) = n \text{ i.e., } V(\bar{X}) = \frac{1}{na^2}$$

$$\text{Hence standard error (S.E.) of } \bar{X} = \sqrt{V(\bar{X})} = \frac{1}{a\sqrt{n}}$$

**Example 8.37.** Let  $X \sim \beta_1(\mu, v)$  and  $Y \sim \gamma(\lambda, \mu + v)$  be independent random variables, ( $\mu, v, \lambda > 0$ ). Find a p.d.f. for  $XY$  and identify its distribution.

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

**Solution.** Since  $X$  and  $Y$  are independently distributed, their joint p.d.f. is given by

$$f(x, y) = \frac{1}{B(\mu, v)} \cdot x^{\mu-1} (1-x)^{v-1} \times \frac{\lambda^{\mu+v}}{\Gamma(\mu+v)} e^{-\lambda y} y^{\mu+v-1};$$

$$0 < x < 1, 0 < y < \infty$$

Let us transform to the new variables  $U$  and  $Z$  by the transformation

$$xy = u, x = z \text{ i.e., } x = z \text{ and } y = u/z$$

Jacobian of transformation  $J$  is given by

$$J = \frac{\partial(x, y)}{\partial(u, z)} = \begin{vmatrix} 0 & 1 \\ \frac{1}{z} & \frac{-u}{z^2} \end{vmatrix} = -\frac{1}{z}$$

Thus the joint p.d.f. of  $U$  and  $Z$  becomes

$$g(u, z) = \frac{\lambda^{\mu+v}}{B(\mu, v) \Gamma(\mu+v)} \cdot (z)^{\mu-1} (1-z)^{v-1} e^{-\lambda u/z} \left( \frac{u}{z} \right)^{\mu+v-1} |J|;$$

$$0 < u < \infty, 0 < z < 1$$

Integrating w.r.t.  $z$  in the range  $0 < z < 1$ , the marginal p.d.f. of  $U$  is given by

$$= \frac{\lambda^{\mu+v} u^{\mu+v-1}}{\Gamma(\mu) \Gamma(v)} \int_0^1 \frac{(1-z)^{v-1} e^{-\lambda u/z}}{z^{v+1}} dz$$

$$= \frac{\lambda^{\mu+v} u^{\mu+v-1}}{\Gamma(\mu) \Gamma(v)} \int_0^1 \frac{1}{z^2} \left( \frac{1}{z} - 1 \right)^{v-1} e^{-\lambda u/z} dz$$

Thus

$$\begin{aligned}
 g(u) &= \frac{\lambda^{\mu+\nu} u^{\mu+\nu-1}}{\Gamma(\mu) \Gamma(\nu)} \int_0^\infty t^{\nu-1} e^{-\lambda u(1+t)} (-dt) \\
 &= \frac{\lambda^{\mu+\nu} u^{\mu+\nu-1} e^{-\lambda u}}{\Gamma(\mu) \Gamma(\nu)} \int_0^\infty e^{-\lambda u t} t^{\nu-1} dt \\
 &= \frac{\lambda^{\mu+\nu} u^{\mu+\nu-1} e^{-\lambda u}}{\Gamma(\mu) \Gamma(\nu)} \frac{\Gamma(\nu)}{(\lambda u)^\nu} \\
 &= \frac{\lambda^\mu}{\Gamma(\mu)} \cdot e^{-\lambda u} u^{\mu-1}, \quad 0 < u < \infty
 \end{aligned}$$

Hence  $U = XY$  is distributed as a gamma variate with parameters  $\lambda$  and  $\mu$ , i.e.,  $XY \sim \gamma(\lambda, \mu)$ .

**Example 8-38.** Let  $p \sim \beta_1(a, b)$  where  $a$  and  $b$  are positive integers. After one observes  $p$ , one secures a coin for which the probability of heads is  $p$ . This coin is flipped  $n$  times. Let  $X$  denote the number of heads which result. Find  $P(X = k)$  for  $k = 0, 1, 2, \dots, n$ . Express the answer in terms of binomial co-efficients.

**Solution.** Since  $p \sim \beta_1(a, b)$ , its p.d.f. is given by

$$P(p) = \frac{1}{B(a, b)} \cdot p^{a-1} (1-p)^{b-1}, \quad 0 < p < 1$$

$P(X = k \mid \text{the probability of success in a single trial is } p)$

$$\begin{aligned}
 P(X = k) &= \int_0^1 P(p) P(X = k \mid p) dp \\
 &= \int_0^1 \frac{1}{B(a, b)} \cdot p^{a-1} (1-p)^{b-1} \cdot \binom{n}{k} p^k (1-p)^{n-k} dp \\
 &= \frac{\binom{n}{k}}{B(a, b)} \int_0^1 p^{a+k-1} (1-p)^{n+b-k-1} dp \\
 &= \frac{\binom{n}{k} B(a+k, n+b-k)}{B(a, b)} \quad \dots(1)
 \end{aligned}$$

We have

$$\frac{1}{B(m, n)} = \frac{\Gamma(m+n)}{\Gamma(m) \Gamma(n)} = \frac{(m+n-1)!}{(m-1)! (n-1)!} = \frac{mn}{m+n} \binom{m+n}{m} \quad \dots(2)$$

$$\therefore P(X = k) = \frac{\binom{n}{k} \frac{ab}{a+b} \binom{a+b}{a}}{\frac{(a+k)(n+b-k)}{(n+a+b)} \binom{n+a+b}{a+k}}$$

$$= \frac{\binom{n}{k} \binom{a+b}{a}}{\binom{n+a+b}{a+k}} \cdot \frac{ab(n+a+b)}{(a+b)(a+k)(n+b-k)}$$

**Example 8.39.** Given the Incomplete Beta Function,

$$B_x(l, m) = \int_0^x t^{l-1} (1-t)^{m-1} dt$$

and  $I_x(l, m) = B_x(l, m)/B(l, m)$ , show that

$$I_x(l, m) = 1 - I_{1-x}(m, l).$$

**Solution.** We have

$$\begin{aligned} I_x(l, m) B(l, m) &= B_x(l, m) = \int_0^x t^{l-1} (1-t)^{m-1} dt \\ &= \int_0^1 t^{l-1} (1-t)^{m-1} dt - \int_x^1 t^{l-1} (1-t)^{m-1} dt \\ &= B(l, m) - \int_x^1 t^{l-1} (1-t)^{m-1} dt \end{aligned} \quad \dots (*)$$

In the integral, put  $1-x = y$ , then

$$\begin{aligned} I_x(l, m) . B(l, m) &= B(l, m) - \int_{1-x}^1 (1-y)^{l-1} y^{m-1} (-dy) \\ &= B(l, m) - \int_0^{1-x} y^{m-1} (1-y)^{l-1} dy \\ &= B(l, m) - B_{1-x}(m, l) = B(l, m) - I_{1-x}(m, l) B(m, l) \end{aligned} \quad [\text{From } (*)]$$

Since  $B(l, m) = B(m, l)$ , we get on dividing throughout by  $B(l, m)$

$$I_x(l, m) = 1 - I_{1-x}(m, l)$$

### EXERCISE 8(d)

1. (a) Suppose the frequency function of a random variable is given by

$$f(x) = \begin{cases} \frac{x^k e^{-x}}{k!}, & \text{for } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

where  $k$  is non-negative integer.

(i) Find the moment generating function of this distribution.

(ii) Determine the mean and variance of this distribution. using moment generating function.

(b) If  $X$  is a Gamma variate with parameter  $\lambda$ , obtain its m.g.f. Hence deduce that the m.g.f. of standard gamma variate tends to  $e^{x^2/2}$  as  $\lambda \rightarrow \infty$ . Also interpret the result. [Delhi Univ. B.A. (Stat. Hons.), 1988, '82]

(c)  $X$  and  $Y$  are two independent gamma variates, with parameters  $l$  and  $m$ . Prove that  $(X + Y)$  is a gamma variate with parameter  $(l + m)$ .

(d) If  $X_1, X_2, \dots, X_n$  are independent and identically distributed gamma random variables, what is the distribution of  $X_1 + X_2 + \dots + X_n$  ?

[Delhi Univ. B. Sc. (Maths. Hons.), 1988]

(e) Consider a random variable  $X$  with the following p.d.f.

$$f(x) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-x/\beta}; 0 < x < \infty; \alpha, \beta > 0$$

Find the moment generating function of  $X$ .

Let the random variable  $X$  with above p.d.f. be defined as  $X \sim Ga(\alpha, \beta)$ . Then prove the following theorems :

**Theorem 1 :** If  $X$  and  $Y$  are independent and  $X \sim Ga(\alpha_1, \beta)$  and  $Y \sim Ga(\alpha_2, \beta)$ , then  $X + Y \sim Ga(\alpha_1 + \alpha_2, \beta)$

**Theorem 2 :** If  $X$  and  $Y$  are independent and  $X \sim Ga(\alpha_1, \beta)$  and  $X + Y \sim Ga(\alpha_1 + \alpha_2, \beta)$ , then  $Y \sim Ga(\alpha_2, \beta)$  [Mysore Univ. B.Sc. 1993]

2. (a) Define the Beta variate of first kind. Obtain its mean and variance. Also define the Beta variate of second kind and state its relation with Gamma variates. [Nagpur Univ. B.Sc. 1993]

(b) Write down the Beta probability functions of the first kind and the second kind with parameters  $\mu$  and  $v$ . Show that a Beta variate of the first kind can be obtained by a transformation of a Beta variate of the second kind.

(c) If  $X$  has a Beta distribution, can  $E(1/X)$  be unity ?

$$\text{Ans. } X \sim B(m, n); E(1/X) = \frac{m + n - 1}{n - 1} > 1. \text{ No.}$$

3. Let  $X \sim \gamma(\lambda, a)$  and  $Y \sim \gamma(\lambda, b)$ , be independent random variables. Show that :

$$E\left[\frac{X}{X + Y}\right] = \frac{E(X)}{E(X + Y)}$$

**Hint.** Since  $X \sim \gamma(\lambda, a)$  and  $Y \sim \gamma(\lambda, b)$  are independent,  $U = X/(X + Y)$  and  $V = (X + Y)$  are independent. Hence

$$E(UV) = E(U)E(V) \Rightarrow E(X) = E\left[\frac{X}{X + Y}\right] \cdot E(X + Y).$$

4. (a) If  $X \sim \gamma(\lambda, \mu)$  and  $\gamma - \gamma(\lambda, \nu)$  are independent random variables, show that

$$\frac{X}{Y} \sim \beta_2(\mu, \nu)$$

(b)  $X$  and  $Y$  are independent Gamma variates find the distribution of  $X/(X+Y)$ .

(c) If  $X$  is random variable having as its  $r$ th moment

$$\mu'_r = \frac{(k+r)!}{k!}$$

$k$  being a positive integer, show that its probability density function is

$$f(x) = \begin{cases} \frac{x^k}{k!} e^{-x}, & x > 0 \\ 0, & x < 0. \end{cases}$$

(d) If the r.v.  $X$  is such that

$$E(X^n) = (n+k)! k^n / k!; n = 1, 2, 3, \dots$$

$k$  being a positive integer, find the p.d.f. of  $X$ .

Ans.  $X - \gamma\left(\frac{1}{k}, k+1\right)$

[Delhi Univ. M.A. (Econ.), 1987]

5. If  $X$  and  $Y$  are independent Gamma variates with parameters  $\mu$  and  $\nu$  respectively, show that the variables

$$U = X + Y \text{ and } V = \frac{X}{X+Y}$$

are independent variables.

6. If  $X$  and  $Y$  are independent Gamma variates with parameter  $\lambda$  and  $\mu$  respectively, show that the variables:

(a)  $U = X + Y \text{ and } V = \frac{X}{X+Y}$

are independently distributed and identify their distributions.

[Delhi Univ. B.Sc. (Stat Hons.) 1991]

(b)  $U = X + Y$  and  $V = X/Y$  are independently distributed,  $U \sim \gamma(\lambda + \mu)$  and  $V \sim \beta_2(\lambda, \mu)$ .

7. A simple sample of  $n$  values  $x_1, x_2, \dots, x_n$  is drawn from the population:

$$dP(x) = \frac{1}{\Gamma(n)} e^{-x} x^{n-1} dx, 0 \leq x < \infty$$

If  $\bar{x}$  is the mean of the sample, find the distribution of  $n\bar{x}$ . Hence find the mean and variance of the distribution.

8. (a) show that for a  $\gamma(\lambda)$  distribution,

$$\frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{1}{\sqrt{\lambda}} = \frac{1}{2} \frac{\mu_3}{\sigma^3}.$$

Show that the excess of kurtosis of the distribution is  $6/\lambda$ .

(b) Show that the mean value of the positive square root of a  $\gamma(\lambda, n)$  variate is  $\Gamma(n + \frac{1}{2}) / [\sqrt{\lambda} \Gamma(n)]$ .

Hence prove that the mean deviation of a  $N(\mu, \sigma^2)$  variate from its mean is  $\sigma \sqrt{2/\pi}$

[Gauhati Univ. M.A. (Eco.), 1991; Delhi Univ. B.Sc. (Stat Hons.), 1989]

**Hint.** Proceed as in Example 8.31.

9. Show that the mean value of the positive square root of  $\beta(\mu, v)$  variate is

$$\frac{\Gamma\left(\mu + \frac{1}{2}\right)\Gamma(\mu + v)}{\Gamma(\mu)\Gamma\left(\mu + v + \frac{1}{2}\right)}$$

10. (a) For the distribution :

$$dP(x) = \frac{1}{B(\mu, v)} \frac{x^{\mu-1}}{(1+x)^{\mu+v}} ; 0 < x < \infty, v > 2$$

Show that variance is  $\frac{\mu(\mu+v-1)}{(v-1)^2(v-2)}$ .

Find also the mode and  $\mu_r'$  (about origin). Also show that harmonic mean is  $(\mu-1)/v$ .

(b) Find the arithmetic mean, harmonic mean and variance of a Beta distribution of first kind with parameter  $\mu$  and  $v$ . Verify that A.M. > H.M.

Also prove that if  $G$  is the geometric mean, then:

$$\log G = \frac{1}{B(\mu, v)} \frac{\partial}{\partial v} \beta(\mu, v) = \frac{\partial}{\partial v} [\log \sqrt{2} - \log \Gamma(\mu + v)]$$

11. Given the Beta distribution in the following form :

$$p(x) = \frac{1}{B(\alpha+1, \lambda+1)} \cdot x^\alpha (1-x)^\lambda ; \alpha > -1, \lambda > -1, 0 \leq x \leq 1$$

Find its variance.

Also find the distribution of (i)  $\frac{1}{X}$ , (ii)  $\frac{1-X}{X}$ .

12. If  $X$  is a normal variate with mean  $\mu$  and standard deviation  $\sigma$ , find the mean and variance of  $Y$  defined by

$$Y = \frac{1}{2} \left( \frac{X-\mu}{\sigma} \right)^2 \quad (\text{Meerut Univ. B. Sc., 1993})$$

8.6. **The Exponential Distribution.** A continuous random variable  $X$  assuming non-negative values is said to have an exponential distribution with parameter  $\theta > 0$ , if its p.d.f. is given by

$$f(x) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad \dots(8.24)$$

The cumulative distribution function  $F(x)$  is given by

$$F(x) = \int_0^x f(u) du = \theta \int_0^x \exp(-\theta u) du$$

$$F(x) = \begin{cases} 1 - \exp(-\theta x), & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad \dots [8.24(a)]$$

### 8.6.1. Moment Generating Function of Exponential Distribution

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \theta \int_0^\infty e^{tx} e^{-\theta x} dx \\ &= \theta \int_0^\infty \exp\{-(\theta-t)x\} dx = \frac{\theta}{(\theta-t)}, \quad \theta > t \end{aligned}$$

$$= \left(1 - \frac{t}{\theta}\right)^{-1} = \sum_{r=0}^{\infty} \left(\frac{t}{\theta}\right)^r$$

$$\therefore \mu'_r = E(X^r) = \text{Coefficient of } \frac{t^r}{r!} \text{ in } M_X(t) \\ = \frac{r!}{\theta^r}; \quad r = 1, 2, \dots$$

$$\therefore Mean = \mu'_1 = \frac{1}{\theta}$$

$$Variance = \mu'_2 = \mu'_1^2 - \mu'_1^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}$$

**Theorem.** If  $X_1, X_2, \dots, X_n$  are independent random variables,  $X_i$  having an exponential distribution with parameter  $\theta_i; i = 1, 2, \dots, n$ ; then  $Z = \min(X_1, X_2, \dots, X_n)$  has exponential distribution with parameter  $\sum_{i=1}^n \theta_i$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1986]

$$\begin{aligned} \text{Proof. } G_Z(z) &= P(Z \leq z) = 1 - P(Z > z) \\ &= 1 - P[\min(X_1, X_2, \dots, X_n) > z] \\ &= 1 - P[X_i > z, \quad i = 1, 2, \dots, n] \\ &= 1 - \prod_{i=1}^n P(X_i > z) = 1 - \prod_{i=1}^n [1 - P(X_i \leq z)] \\ &= 1 - \prod_{i=1}^n [1 - F_{X_i}(z)] \end{aligned}$$

where  $F$  is the distribution function of  $X_i$ .

$$= 1 - \prod_{i=1}^n \left[ 1 - \left(1 - e^{-\theta_i z}\right) \right]$$

[c.f. 8.24(a)]

$$= \begin{cases} 1 - \exp \left\{ \left( - \sum_{i=1}^n \theta_i \right) z \right\}, & z > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\therefore g_Z(z) = \begin{cases} \left( \sum_{i=1}^n \theta_i \right) \exp \left\{ \left( - \sum_{i=1}^n \theta_i \right) z \right\}, & z > 0 \\ 0, & \text{otherwise} \end{cases}$$

$\Rightarrow Z = \min(X_1, X_2, \dots, X_n)$  is an exponential variate with parameter  $\sum_{i=1}^n \theta_i$ .

**Cor.** If  $X_i ; i = 1, 2, \dots, n$  are identically distributed, following exponential distribution with parameter  $\theta$ , then  $Z = \min(X_1, X_2, \dots, X_n)$  is also exponentially distributed with parameter  $n\theta$ .

**Example 8.40.** Show that the exponential distribution "lacks memory", i.e., if  $X$  has an exponential distribution, then for every constant  $a \geq 0$ , one has  $P(Y \leq x | X \geq a) = P(Y \leq x)$  for all  $x$ , where  $Y = X - a$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1989; Calicut Univ. B.Sc. (Main Stat.), 1991]

**Solution.** The p.d.f. of the exponential distribution with parameter  $\theta$  is

$$f(x) = \theta e^{-\theta x}; \theta > 0, 0 < x < \infty$$

We have

$$\begin{aligned} P(Y \leq x \cap X \geq a) &= P(X - a \leq x \cap X \geq a) && (\because Y = X - a) \\ &= P(X \leq a + x \cap X \geq a) = P(a \leq X \leq a + x) \\ &= \theta \int_a^{a+x} e^{-\theta x} dx = e^{-a\theta} (1 - e^{-\theta x}) \end{aligned}$$

and

$$P(X \geq a) = \theta \int_a^\infty e^{-\theta x} dx = e^{-a\theta}$$

$$\therefore P(Y \leq x | X \geq a) = \frac{P(Y \leq x \cap X \geq a)}{P(X \geq a)} = 1 - e^{-\theta x} \quad \dots(*)$$

Also  $P(X \leq x) = \theta \int_0^x e^{-\theta x} dx = 1 - e^{-\theta x} \quad \dots(**)$

From (\*) and (\*\*), we get

$$P(Y \leq x | X \geq a) = P(Y \leq x)$$

i.e., exponential distribution lacks memory.

**Example 8.41.**  $X$  and  $Y$  are independent with a common p.d.f. (exponential):

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Find a p.d.f. for  $X - Y$ . [Delhi Univ. B.Sc. (Stat. Hons.), 1988, '85]

**Solution.** Since  $X$  and  $Y$  are independent and identically distributed (i.i.d.), their joint p.d.f. is given by

$$f_{XY}(x, y) = \begin{cases} e^{-(x+y)}, & x > 0, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Let } \begin{cases} u = x - y \\ v = y \end{cases} \Rightarrow \begin{cases} x = u + v \\ y = v \end{cases} \quad \dots(1)$$

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1$$

Thus the joint p.d.f. of  $U$  and  $V$  becomes

$$g(u, v) = e^{-(u+2v)}; v > 0, -\infty < u < \infty$$

$$(1) \Rightarrow u = x - v \Rightarrow v = x - u$$

Thus  $v > -u$  if  $-\infty < u < 0$   
and  $v > 0$  if  $u > 0$   
For  $-\infty < u < 0$ ,

$$g(u) = \int_{-u}^{\infty} g(u, v) dv = \int_{-u}^{\infty} e^{-(u+2v)} dv = e^{-u} \left| \frac{e^{-2v}}{-2} \right|_{-u}^{\infty} = \frac{1}{2} e^u$$

and for  $u > 0$ ,

$$g(u) = \int_{-u}^{\infty} g(u, v) dv = e^{-u} \left| \frac{e^{-v}}{-2} \right|_{-u}^{\infty} = \frac{1}{2} e^{-u^2}$$

Hence the p.d.f of  $U = X - Y$  is given by

$$g(u) = \begin{cases} \frac{1}{2} e^u, & -\infty < u < 0 \\ \frac{1}{2} e^{-u}, & u > 0 \end{cases}$$

These results can be combined to give

$$g(u) = \frac{1}{2} e^{-|u|}, -\infty < u < \infty$$

which is the p.d.f. of standard Laplace distribution (c.f. § 8.7).

Aliter.

$$M_X(t) = \int_0^{\infty} e^{tx} f(x) dx = \int_0^{\infty} e^{-(1-t)x} dx = \left| \frac{e^{-(1-t)x}}{-(1-t)} \right|_0^{\infty} = \frac{1}{1-t}, t < 1$$

∴ Characteristic function of  $X$  is

$$\varphi_X(t) = \frac{1}{1-it} = \varphi_Y(t),$$

(since  $X$  and  $Y$  are identically distributed.)

$$\therefore \varphi_{X-Y}(t) = \varphi_{X+(-Y)}(t) = \varphi_{X(t)} \varphi_{-Y(t)} \quad (\because X, Y \text{ are independent})$$

$$= \varphi_X(t) \cdot \varphi_Y(-t) = \frac{1}{(1-it)(1+it)} = \frac{1}{1+t^2}$$

which is the characteristic function of the Laplace distribution, (c.f. § 8-7)

$$g(u) = \frac{1}{2} e^{-|u|}, -\infty < u < \infty \quad \dots (*)$$

Hence by the uniqueness theorem of characteristic functions,  $U = X - Y$  has the p.d.f. given in (\*).

**8-7. Laplace (Double Exponential) Distribution.** A continuous random variable  $X$  is said to follow standard Laplace distribution if its p.d.f. is given by

$$f(x) = \frac{1}{2} e^{-|x|}, -\infty < x < \infty \quad \dots (8-25)$$

Characteristic function is given by

$$\begin{aligned} \varphi_X(t) &= \int_{-\infty}^{\infty} e^{itx} f(x) dx = \frac{1}{2} \int_{-\infty}^{\infty} e^{itx} \cdot e^{-|x|} dx \\ &= \frac{1}{2} \left[ \int_{-\infty}^{\infty} \cos tx \cdot e^{-|x|} dx + i \int_{-\infty}^{\infty} \sin tx \cdot e^{-|x|} dx \right] \\ &= \frac{1}{2} \cdot 2 \int_0^{\infty} \cos tx \cdot e^{-|x|} dx, \end{aligned}$$

Since the integrands in the first and second integrals are even and odd function of  $x$  respectively.

$$\begin{aligned} \therefore \varphi_X(t) &= \int_0^{\infty} e^{-x} \cos tx dx \\ &= 1 - t^2 \int_0^{\infty} e^{-x} \cos tx dx \quad (\text{on integration by parts}) \\ &= 1 - t^2 \varphi_X(t) \\ \Rightarrow \varphi_X(t) &= \frac{1}{1+t^2} \quad \dots (8-25 a) \end{aligned}$$

The mean of this distribution is zero, standard deviation is  $\sqrt{2}$  and mean deviation about mean is 1.

**Remark. Generalised Laplace Distribution.** A continuous r.v.  $X$  is said to have Laplace distribution with two parameters  $\lambda$  and  $\mu$  if its p.d.f. is given by

$$f(x) = \frac{1}{2\lambda} \exp[-|x-\mu|\lambda], -\infty < x < \infty; \lambda > 0 \quad \dots (8-26)$$

Taking  $U = \frac{X-\mu}{\lambda}$ , in (8-26) we obtain the p.d.f. of standard Laplace variate given in (8-25).

**Moments.** The  $r$ th moment about origin is given by

$$\begin{aligned}
 \mu'_r &= E(X^r) = \frac{1}{2\lambda} \int_{-\infty}^{\infty} x^r \exp\left(\frac{-|x-\mu|}{\lambda}\right) dx \\
 &= \frac{1}{2} \int_{-\infty}^{\infty} (z\lambda + \mu)^r \exp(-|z|) dz, \quad \left[ z = \frac{x-\mu}{\lambda} \right] \\
 &= \frac{1}{2} \int_{-\infty}^{\infty} \left[ \sum_{k=0}^r \binom{r}{k} (z\lambda)^k \mu^{r-k} \right] \exp(-|z|) dz \\
 &= \frac{1}{2} \sum_{k=0}^r \left[ \binom{r}{k} \lambda^k \mu^{r-k} \int_{-\infty}^{\infty} z^k \exp(-|z|) dz \right] \\
 &= \frac{1}{2} \sum_{k=0}^r \left[ \binom{r}{k} \lambda^k \mu^{r-k} \left\{ \int_{-\infty}^0 z^k e^{-|z|} dz + \int_0^{\infty} z^k e^{-|z|} dz \right\} \right] \\
 &= \frac{1}{2} \sum_{k=0}^r \left[ \binom{r}{k} \lambda^k \mu^{r-k} \left\{ (-1)^k \int_0^{\infty} e^{-z} z^k dz + \int_0^{\infty} e^{-z} z^k dz \right\} \right] \\
 &= \frac{1}{2} \sum_{k=0}^r \left[ \binom{r}{k} \lambda^k \mu^{r-k} \Gamma(k+1) \{(-1)^k + 1\} \right]
 \end{aligned}$$

$$\Rightarrow \mu'_r = \frac{1}{2} \sum_{k=0}^r \left[ \binom{r}{k} \lambda^k \mu^{r-k} k! \{1 + (-1)^k\} \right] \quad \dots(8.26a)$$

$$\therefore \text{Mean} = \mu'_r = \mu_1' = \mu \text{ and } \mu_2' = \mu^2 + 2\lambda^2$$

$$\therefore \sigma_X^2 = \mu_2' - \mu_1'^2 + 2\lambda^2.$$

Similarly we can obtain higher order moments from (8.26 a) and hence the values of  $\beta_1$  and  $\beta_2$  can be obtained.

The characteristic function of (8.26) can be obtained exactly similarly as we obtained the characteristic function of standard Cauchy distribution, c.f. § 8.9.

**8.8. Weibul Distribution.** A random variable  $X$  has a Weibul distribution with three parameters  $c (> 0)$ ,  $\alpha (> 0)$  and  $\mu$  if the r.v.

$$Y = \left( \frac{X-\mu}{\alpha} \right)^c \quad \dots(i)$$

has the exponential distribution with p.d.f.

$$p_Y(y) = e^{-y}, y > 0 \quad \dots(ii)$$

The p.d.f. of  $X$  is given by

$$p_X(x) = c \alpha^{-1} \left( \frac{x - \mu}{\alpha} \right)^{-1} \exp \left[ - \left( \frac{x - \mu}{\alpha} \right)^c \right], \quad x > \mu, c > 0 \quad \dots(iii)$$

The standard Weibul distribution is obtained on taking  $\alpha = 1$  and  $\mu = 0$ , so that the p.d.f. of *standard Weibul distribution* which depends only on a single parameter  $c$  is given by

$$p_X(x) \doteq c x^{c-1} \cdot \exp(-x^c); \quad x > 0, c > 0 \quad \dots(iv)$$

### 8.8.1. Moments of Standard Weibul Distribution (iv)

For standard Weibul distribution, ( $\alpha = 1, \mu = 0$ ), from (i), we get  $Y = X^c$  which has the exponential distribution (ii). We have

$$\begin{aligned} \mu'_r &= E(X^r) = E(Y^{r/c})' = E(Y^{r/c}) \\ &= \int_0^\infty e^{-y} \cdot y^{r/c} dy \\ \Rightarrow \mu'_r &= \Gamma\left(\frac{r}{c} + 1\right) \quad [\because Y \text{ has p.d.f. (ii)}] \\ \therefore \text{Mean} &= E(X) = \Gamma\left(\frac{1}{c} + 1\right) \\ \text{and } \text{Var}(X) &= E(X^2) = [E(X)]^2 \\ &= \Gamma\left(\frac{2}{c} + 1\right) - \left[ \Gamma\left(\frac{1}{c} + 1\right) \right]^2 \end{aligned} \quad \dots(v)$$

Similarly, we can obtain expressions for higher order moments and hence for  $\beta_1$  and  $\beta_2$ . For large  $c$ , the mean is approximated by

$$\begin{aligned} E(X) &\approx 1 - \frac{\gamma}{c} + \frac{1}{2c^2} \left( \frac{\pi^2}{6} + \gamma^2 \right) \\ &= 1 - 0.57722 c^{-1} + 0.98905 c^{-2} \end{aligned}$$

where  $\gamma = 0.57722$  is Euler's constant.

The distribution is named after Waloddi Weibul, a Swedish physicist, who used it in 1939 to represent the distribution of the breaking strength of materials. Kao, J.H.K. (1958-59) advocated the use of this distribution in reliability studies and quality control work. It is also used as a tolerance distribution in the analysis of quantum response data.

**8.8.2. Characterisation of Weibul Distribution.** Dubey, S.D. (1968) has obtained the following result:

"Let  $X_i$  ( $i = 1, 2, \dots, n$ ) be i.i.d. random variables. Then  $\min(X_1, X_2, \dots, X_n)$  has a Weibul distribution if and only if the common distribution of  $X_i$ 's is a Weibul distribution".

**Proof.** Let  $X_i$ , ( $i = 1, 2, \dots, n$ ) be i.i.d. r.v. each with Weibul distribution (iii) and let  $Y = \min(X_1, X_2, \dots, X_n)$ . Then

$$P_r(Y > y) = P[\min(X_1, X_2, \dots, X_n) > y]$$

$$\begin{aligned}
 &= P \left[ \bigcap_{i=1}^n X_i > y \right] \\
 &= \prod_{i=1}^n P(X_i > y) = \left[ P(X_i > y) \right]^n \quad \dots(*) 
 \end{aligned}$$

since  $X_i$ 's are i.i.d. r.v.'s.

$$\begin{aligned}
 \text{Now } P(X_i > y) &= \int_y^\infty c \alpha^{-1} \left( \frac{x-\mu}{\alpha} \right)^{c-1} \cdot \exp \left[ -\left( \frac{x-\mu}{\alpha} \right)^c \right] dx \\
 &= \int_{\left( \frac{y-\mu}{\alpha} \right)^c}^\infty e^{-t} dt \quad \left[ t = \left( \frac{x-\mu}{\alpha} \right)^c \right] \\
 &= \exp \left[ -\left( \frac{y-\mu}{\alpha} \right)^c \right].
 \end{aligned}$$

Substituting in (\*), we get

$$\begin{aligned}
 P(Y > y) &= \left[ \exp \left\{ -\left( \frac{y-\mu}{\alpha} \right)^c \right\} \right]^n \\
 &= \exp \left[ -n \left( \frac{y-\mu}{\alpha} \right)^c \right] \\
 &= \exp \left[ - \left\{ \frac{n^{1/c}(y-\mu)}{\alpha} \right\}^c \right]
 \end{aligned}$$

This implies that  $Y$  has the same Weibul distribution as  $X_i$ 's with the difference that the parameter  $\alpha$  is replaced by  $\alpha n^{-1/c}$ .

**8.8.3. Logistic Distribution.** A continuous r.v.  $X$  is said to have a Logistic distribution with parameters  $\alpha$  and  $\beta$ , if its distribution function is of the form:

$$F_X(x) = [1 + \exp \{-(x-\alpha)/\beta\}]^{-1}, \beta > 0 \quad \dots(8.26 b)$$

$$= \frac{1}{2} \left[ 1 + \tanh \left\{ \frac{1}{2} (x-\alpha)/\beta \right\} \right]; \beta > 0 \quad \dots(8.26 c)$$

(See Remark 1 on page 8.94).

The p.d.f. of Logistic distribution with parameters  $\alpha$  and  $\beta (> 0)$  is given by

$$f(x) = \frac{d}{dx} (F(x))$$

$$= \frac{1}{\beta} [1 + \exp \{-(x-\alpha)/\beta\}]^{-2} [\exp \{-(x-\alpha)/\beta\}] \quad \dots(8.26 d)$$

$$= \frac{1}{4\beta} \operatorname{sech}^2 \left\{ \frac{1}{2} (x-\alpha)/\beta \right\} \quad \dots(8.26 e)$$

The p.d.f. of standard Logistic variate  $Y = (X-\alpha)/\beta$ , is given by:

$$g_Y(y) = f(x) \cdot \left| \frac{dx}{dy} \right| \quad \dots(8.26f)$$

$$= e^{-y} \left( 1 + e^{-y} \right)^{-2}; -\infty < y < \infty \quad \dots(8.26f)$$

$$= \frac{1}{4} \operatorname{sech}^2 \left( \frac{1}{2} y \right); -\infty < y < \infty \quad \dots(8.26g)$$

The distribution function of  $Y$  is :

$$G_Y(y) = (1 + e^{-y})^{-1}; -\infty < y < \infty \quad \dots(8.26h)$$

Logistic distribution is extensively used as growth function in population and demographic studies and in time series analysis. Theoretically, Logistic distribution can be obtained as :

(i) The limiting distribution (as  $n \rightarrow \infty$ ) of the standardised mid range, (average of the smallest and the largest sample observations), in random samples of size  $n$

(ii) A mixture of extreme value distributions.

**Moment Generating Function.** The m.g.f. of standard Logistic variate  $Y$  is given by:

$$\begin{aligned} \mu_Y(t) &= E(e^{tY}) = \int_{-\infty}^{\infty} e^{ty} \cdot g(y) dy \\ &= \int_{-\infty}^{\infty} e^{ty} \cdot e^{-y} (1 + e^{-y})^{-2} dy \\ &= \int_{-\infty}^{\infty} e^{ty} e^{-y} \left( \frac{1 + e^y}{e^y} \right)^{-2} dy \\ &= \int_{-\infty}^{\infty} e^{ty} e^y (1 + e^y)^{-2} dy \end{aligned}$$

$$\text{Put } z = (1 + e^y)^{-1} \Rightarrow e^y = \frac{1}{z} - 1 = \frac{1-z}{z}$$

$$\begin{aligned} \therefore M_Y(t) &= \int_1^0 \left( \frac{1-z}{z} \right)' \cdot (-dz) = \int_0^1 z^{-t} (1-z)^t dz \\ &= \beta(1-t, 1+t), \quad 1-t > 0 \\ &= \Gamma(1-t) \Gamma(1+t) / \Gamma 2 \\ &= \Gamma(1-t) \Gamma(1+t) \\ &= \pi t \operatorname{cosec} \pi t; t < 1 \quad \dots(8.26i) \\ &= 1 + \frac{\pi^2 t^2}{6} + \frac{7}{360} \pi^4 t^4 + \dots(*) \end{aligned}$$

(See Remark 2 below.)

$$\begin{aligned} \therefore E(Y) &= \text{Coefficient of } t \text{ in } (*) = 0 \\ &\Rightarrow \text{Mean} = 0 \end{aligned}$$

$$\therefore \mu_2 = E(Y^2) = \text{Coefficient of } \frac{t^2}{2!} \text{ in } (*) = \frac{\pi^2}{3} ,$$

$$\mu_3 = E(Y^3) = 0$$

$$\mu_4 = E(Y^4) = \text{Coefficient of } \frac{t^4}{4!} \text{ in } (*) = \frac{7}{15} \pi^4$$

Hence for standard Logistic distribution :

$$\text{Mean} = 0, \text{ Variance} = \mu_2 = \pi^2/3,$$

$$\beta_1 = \frac{\mu_3}{\mu_2^2} = 0, \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{7 \times 9}{15} = 4.2$$

**Remarks 1.** We have :

$$\begin{aligned} \tanh x &= \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}} \\ \Rightarrow 1 + \tanh x &= \frac{2}{1 + e^{-2x}} \Rightarrow \frac{1}{2} [1 + \tanh x] = (1 + e^{-2x})^{-1} \end{aligned}$$

$$2. \quad x \operatorname{cosec} x = 1 + \frac{x^2}{6} + \frac{7}{360} x^4 + \dots$$

$$\begin{aligned} \text{Proof. } x \operatorname{cosec} x &= \frac{x}{\sin x} = \frac{x}{x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots} \\ &= \left[ 1 - \left( \frac{x^2}{6} - \frac{x^4}{120} + \frac{x^6}{720} \dots \right) \right]^{-1} \\ &= 1 + \left( \frac{x^2}{6} - \frac{x^4}{120} + \dots \right) + \left( \frac{x^2}{6} - \frac{x^4}{120} + \dots \right)^2 + \dots \\ &= 1 + \frac{x^2}{6} + x^4 \left( \frac{1}{36} - \frac{1}{120} \right) + \dots \\ &= 1 + \frac{x^2}{6} + \frac{7}{360} x^4 + \dots \end{aligned}$$

3. We have:

$$g(y) = e^{-y} \left( 1 + \frac{1}{e^y} \right)^{-2} = e^y \left( 1 + e^y \right)^{-2} = g(-y)$$

$\Rightarrow$  The probability curve of  $Y$  is symmetric about the line  $y = 0$ .

Since p.d.f.  $g(y)$  is symmetric about origin ( $y = 0$ ), all odd order moments about origin are zero i.e.,

$$\mu'_{2r+1} = E(Y^{2r+1}) = 0, \quad r = 0, 1, 2, \dots$$

In particular

$$\text{Mean} = \mu'_1 = 0$$

$$\therefore \mu'_r = r\text{th moment about origin}$$

$$\Rightarrow \mu_{2n+1} = \mu'_{2n+1} = 0$$

i.e., all odd order moments about mean of the standard logistic distribution are zero.

In particular  $\mu_3 = 0 \Rightarrow \beta_1 = 0$

4. The mean and variance of the logistic Variable ( $X$ ) with parameters  $\alpha$  and  $\beta$ , are given by:

$$\begin{aligned} E(X) &= E(\alpha + \beta Y) \\ &= \alpha + \beta E(Y) \\ &= \alpha \end{aligned}$$

$$\text{Var } X \equiv \text{Var}(\alpha + \beta Y) = \beta^2 \text{Var}(Y) = \beta^2 \pi^2/3.$$

5. We have :

$$G(y) = (1 + e^{-y})^{-1} = \left( \frac{1 + e^y}{e^y} \right)^{-1} = \frac{e^y}{1 + e^y}$$

$$\Rightarrow 1 - G(y) = 1 - \frac{e^y}{1 + e^y} = \frac{1}{1 + e^y}$$

$$\therefore G(y) \cdot [1 - G(y)] = \frac{e^y}{(1 + e^y)^2} = g(y) \quad \dots(826j)$$

(c.f. Remark 3)

$$\text{Also } \frac{G(y)}{1-G(y)} = e^y \Rightarrow y = \log_e \left[ \frac{G(y)}{1-G(y)} \right] \quad \dots(8.26k)$$

6. Mean deviation for the standard Logistic distribution is

$$2 \left[ 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots \right] = 2 \sum_{i=1}^{\infty} \left[ \frac{(-1)^{i-1}}{i} \right] = 2 \log_e 2$$

**Proof** is left as an exercise to the reader.

### **EXERCISE 8(e)**

1. (a) Show that for the exponential distribution

$$p(x) = y_0 \cdot e^{-x/\sigma}, \quad 0 \leq x < \infty; \quad \sigma > 0,$$

mean and variance are equal. Also obtain the interquartile range of the distribution.

[Delhi Univ. B.Sc. (Stat. Hons.), 1985, 1982]

(b) Suppose that during rainy season on a tropical island the length of the shower has an exponential distribution, with parameter  $\lambda = 2$ , time being measured in minutes. What is the probability that a shower will last more than three minutes? If a shower has already lasted for 2 minutes, what is the probability that it will last for at least one more minute? [Madras Univ. B.Sc. (Main Stat) 1988]

2. (a) If  $X_1, X_2, \dots, X_n$  are independent random variables having exponential distribution with parameter  $\lambda$ , obtain the distribution of  $Y = \sum_{i=1}^n X_i$ .

(b) Obtain the moment generating function and the cumulant generating function of the distribution with p.d.f.

$$f(x) = \frac{1}{\sigma} e^{-x/\sigma}; 0 < x < \infty, \sigma > 0$$

[Madras Univ. B.Sc. (Main Stat.) Oct. 1992]

Hence or otherwise obtain the values of the constants  $\beta_1, \beta_2, \gamma_1$  and  $\gamma_2$ .

(c) A continuous random variable  $X$  has the probability density function  $f(x)$  given by

$$f(x) = A e^{-x^2/2} \quad x > 0 \\ = 0, \quad \text{otherwise}$$

Find the value of  $A$  and show that for any two positive numbers  $s$  and  $t$ ,

$$P[X > s + t | X > s] = P[X > t].$$

3. If  $X_1$  and  $X_2$  are independent and identically distributed each with frequency function  $e^{-x}, x > 0$ , find the frequency function of  $X_1 + X_2$ .

(b) If  $X_1, X_2, \dots, X_n$  are independent r.v.'s  $X_i$  having an exponential distribution with parameter  $\theta_i, (i = 1, 2, \dots, n)$ , then prove that  $Z = \min(X_1, X_2, \dots, X_n)$  has an exponential distribution with parameter  $\sum_{i=1}^n \theta_i$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1990, '88, '86]

4. Let  $X$  and  $Y$  have common p.d.f.  $\alpha e^{-\alpha x}, 0 < x < \infty, \alpha > 0$ . Find the p.d.f. of

(i)  $X^3$ , (ii)  $3 + 2X$ , (iii)  $X - Y$ , and (iv)  $|X - Y|$

Ans. (i)  $\frac{\alpha}{3} x^{-2/3} \exp(-\alpha x^{1/3})$ , (ii)  $\frac{\alpha}{2} e^{-\alpha(x-3)/2}, x > 3$

(iii)  $\frac{\alpha}{2} e^{-\alpha|x|}$ , all  $x$ , and (iv)  $\alpha e^{-\alpha|x|}, x > 0$ .

5. (a)  $X$  and  $Y$  are independent random variables each exponentially distributed with the same parameter  $\theta$ . find p.d.f. for  $\frac{X}{X+Y}$  and identify its distribution.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

(b) The density functions of the independent random variables  $X$  and  $Y$  are:

$$\begin{array}{lll} f_X(x) = \lambda e^{-\lambda x}, & x > 0, \lambda > 0 & f_Y(y) = \lambda e^{-\lambda y}, \quad y > 0, \lambda > 0 \\ = 0 & , x \leq 0 & = 0 & , \text{ otherwise} \end{array}$$

Find the density function of the random variable  $Z = X/Y$ .

6. (a) For the distribution given by the density function

$$f(x) = \frac{1}{2} e^{-|x|}, -\infty < x < \infty,$$

obtain the moment generating function.

(b) Find the characteristic function of standard Laplace distribution and hence find its mean and standard deviation.

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

7. (a) If  $X$  has exponential distribution with mean 2, find  $P(X < 1) | X < 2)$

Ans.  $P(X < 1) [P(X < 2) = (1 - e^{-\theta})/(1 - e^{2\theta}) \text{ where } \theta = 1/2]$ .

[Delhi Univ. B.A. (Spl. Hons. Course-Statistics), 1989]

(b) If  $X \sim \text{Expo}(\lambda)$  with  $P(X \leq 1) = P(X > 1)$ ,

...(\*)

find  $\text{Var } X$ .

[Delhi Univ. B.Sc. (Maths Hons.), 1985]

**Hint.**  $P(X \leq 1) + P(X > 1) = 1 \Rightarrow P(X \leq 1) = 1/2$

[Using (\*)]

Ans.  $\text{Var}(X) = 1/\lambda^2 = 1/(\log e^2)^2$

8. (a) Show that  $Y = -(1/\lambda) \log F(X)$  is  $\text{Expo}(\lambda)$ .

[Delhi Univ. B.A. Hons. (Spl. Course-Statistics), 1985]

$$\text{Hint. } \mu_Y(t) = E(e^{tY}) = E \exp \left[ -\frac{t}{\lambda} \log F(x) \right]$$

$$= E[F(X)^{-t/\lambda}] = E[Z^{-t/\lambda}] \text{ where } Z = F(X) \sim U[0, 1]$$

(b) If  $X_1, X_2, X_3$  and  $X_4$  are i.i.d.  $N(0, 1)$  variates, show that  $Y = X_1 X_2 - X_3 X_4$ , has p.d.f.

$$f(y) = \frac{1}{2} \exp[-|y|], -\infty < y <$$

[Indian Civil Services, 1984]

**Hint.** Show that  $\varphi_Y(t) = 1/(1 + t^2) \Rightarrow Y$  has Standard Laplace distribution.

$$\text{Use : } \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(ax^2 + 2hy + by^2)} dx dy = \frac{\pi}{\sqrt{ab - h^2}}$$

9. 200 electric light bulbs were tested and the average life time of the bulbs was found to be 25 hours. Using the summary given below, test the hypothesis that the lifetime is exponentially distributed.

Lifetime in hours :      0–20      20–40      40–60      60–80      80–100

Number of bulbs :      104      56      24      12      4

[You are given that an exponential distribution with parameters  $\alpha > 0$  has the probability density function:

$$p(x) = \alpha e^{-\alpha x}, (x \geq 0) \\ = 0, \quad (x < 0)$$

[Institute of Actuaries (London), April 1978]

10. Find the first four cumulants of the Laplace distribution defined by

$$f(x) = \frac{1}{2\lambda} \left[ \exp \{-|x - \mu|/\lambda\} \right]; -\infty < x < \infty, \lambda > 0$$

and hence find the values of  $m$ ,  $\sigma$ ,  $\gamma_1$  and  $\gamma_2$ . Calculate also the semi-interquartile range (S.I.R.)

**Ans.**  $\kappa_1 = \mu$ ,  $\kappa_2 = 2\lambda^2$ ,  $\kappa_3 = 0$ ,  $\kappa_4 = 12\lambda^4$ ;  $m = \mu$ ,  $\sigma = \sqrt{2}\lambda$ ,  $\gamma_1 = 0$ ,  $\gamma_2 = 3$  and S.I.R. =  $\lambda \log_e 2$

11. The p.d.f. of a r.v.  $X$  follows the following probability law

$$p(x) = \frac{1}{2\theta} \exp \left( -\frac{|x - \theta|}{\theta} \right), -\infty < x < \infty.$$

Find m.g.f. of  $X$ . Hence or otherwise, find  $E(X)$  and  $\text{Var}(X)$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1986]

12.  $X_i, i = 1, 2, \dots, n$  are i.i.d. r.v.'s having Weibul distribution with three parameters. Show that the variable  $Y = \min(X_1, X_2, \dots, X_n)$ , also has Weibul distribution and identify its parameters.

[Delhi Univ. B.Sc. (Stat. Hons.), 1984]

13. Obtain the moment generating function of Logistic distribution and hence find its mean and variance. [Delhi Univ. B.Sc. (Stat. Hons.), 1993]

14. (a) Obtain the p.d.f.  $g(y)$  and the distribution function  $G(y)$  of the standard logistic variate and prove that :

(i)  $g(y)$  is symmetric about origin.

(ii)  $g(y) = G(y)[1 - G(y)]$

$$(iii) y = \log_e \left[ \frac{G(y)}{1 - G(y)} \right]$$

(b) Obtain the m.g.f. of standard logistic variate and hence prove that:

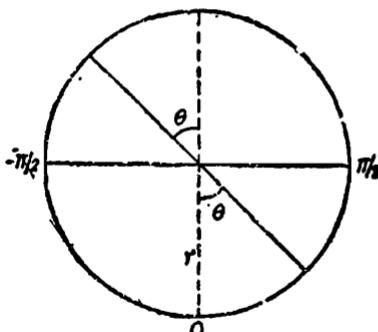
$$\text{Mean} = 0, \quad \text{Variance} = \pi^2/3,$$

$$\beta_1 = 0, \quad \beta_2 = \frac{21}{5}; \quad \mu_{2r+1} = 0$$

and mean deviation about mean =  $2 \log_3 2$ .

**8.9. Cauchy's Distribution.** Let us consider a roulette wheel in which the probability of the pointer stopping at any part of the circumference is constant. In other words, the probability for any value of  $\theta$  lies in the interval  $[-\pi/2, \pi/2]$  is constant and consequently  $\theta$  is a rectangular variate in the range  $[-\pi/2, \pi/2]$  with probability differential given by

$$dP(\theta) = \left\{ \begin{array}{l} (1/\pi) d\theta, \quad -\pi/2 \leq \theta \leq \pi/2 \\ 0, \quad \text{otherwise} \end{array} \right\} \quad \dots(8.27)$$



Let us now transform to the variable  $X$  by the substitution:

$$x = r \tan \theta \Rightarrow dx = r \sec^2 \theta d\theta$$

Since  $-\pi/2 \leq \theta \leq \pi/2$ , the range for  $X$  is from  $-\infty$  to  $\infty$ . Thus the probability differential of  $X$  becomes:

$$dF(x) = \frac{1}{\pi} \cdot \frac{dx}{r \sec^2 \theta} = \frac{1}{\pi} \cdot \frac{dx}{\left| r \right| \left| 1 + (x^2/r^2) \right|} = \frac{r}{\pi} \cdot \frac{dx}{r^2 + x^2}; -\infty < x < \infty$$

In particular if we take  $r = 1$ , we get

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}; -\infty < x < \infty$$

This is the p.d.f. of a standard Cauchy variate and we write  $X \sim C(1,0)$

**Definition.** A random variable  $X$  is said to have a standard Cauchy distribution if its p.d.f. is given by

$$f_X(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty \quad \dots(8.28)$$

and  $X$  is termed as a standard Cauchy variate.

More generally, Cauchy distribution with parameters  $\lambda$  and  $\mu$  has the following p.d.f.,

$$g_Y(y) = \frac{\lambda}{\pi[\lambda^2 + (y - \mu)^2]}, -\infty < y < \infty; \lambda > 0 \quad \dots(8.29)$$

and we write  $X \sim C(\lambda, \mu)$

But putting  $X = (Y - \mu)/\lambda$  in (8.29), we get (8.28).

**8.9.1. Characteristic Function of Cauchy Distribution.** If  $X$  is a standard Cauchy variate then

$$\varphi_X(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{itx}}{1+x^2} dx \quad \dots(*)$$

To evaluate (\*) consider Laplace distribution

$$f_1(z) = \frac{1}{2} e^{-|z|}, -\infty < z < \infty$$

$$\text{Then } \varphi_1(t) = E(e^{itZ}) = \frac{1}{1+t^2} \quad [\text{From (8.25 a)}]$$

Since  $\varphi_1(t)$  is absolutely integrable in  $(-\infty, \infty)$ , we have by Inversion theorem

$$\frac{1}{2} e^{-|z|} = f_1(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} \varphi_1(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itz}}{1+t^2} dt$$

$$\Rightarrow e^{-|z|} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{-itz}}{1+t^2} dt = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{itz}}{1+t^2} dt \quad (\text{Changing } t \text{ to } -t)$$

On interchanging  $t$  and  $z$ , we get

$$e^{-|t|} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{iz}}{1+z^2} dz \quad \dots (**)$$

From (\*) and (\*\*), we get

$$\varphi_X(t) = e^{-|t|} \quad \dots (8.30)$$

**Remarks.** 1. If  $Y$  is a Cauchy variate with parameters  $\lambda$  and  $\mu$ , then

$$X = \frac{Y - \mu}{\lambda} \Rightarrow Y = \mu + \lambda X$$

$$\therefore \varphi_Y(t) = E(e^{itY}) = e^{i\mu t} E(e^{i\lambda X}) = e^{i\mu t} \varphi_X(t\lambda) \\ = e^{i\mu t - \lambda |t|}, \lambda > 0 \quad \dots (8.30a)$$

2. **Additive Property of Cauchy distribution.** If  $X_1$  and  $X_2$  are independent Cauchy variates with parameters  $(\lambda_1, \mu_1)$  and  $(\lambda_2, \mu_2)$  respectively, then  $X_1 + X_2$  is a Cauchy variate with parameters  $(\lambda_1 + \lambda_2, \mu_1 + \mu_2)$ .

**Proof.**  $\varphi_{X_j}(t) = \exp\{i\mu_j t - \lambda_j |t|\}, (j=1, 2)$

$$\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t) \varphi_{X_2}(t) \quad (\text{Since } X_1, X_2 \text{ are independent}) \\ = \exp\left[i t(\mu_1 + \mu_2) - (\lambda_1 + \lambda_2) |t|\right]$$

and the result follows by uniqueness theorem of characteristic functions.

3. Since  $\varphi'_{X_i}(t)$  in (8.30) [where  $'$  denotes differentiation w.r.t.  $t$ ] does not exist at  $t=0$ , the mean of the Cauchy distribution does not exist:

4. Let  $X_1, X_2, \dots, X_n$  be a sample of  $n$  independent observations from a standard Cauchy distribution and define  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . Then

$$\begin{aligned} \varphi_{\bar{X}}(t) &= \varphi_{\sum X_i}(t/n) = \prod_{i=1}^n \{\varphi_{X_i}(t/n)\} \\ &= [\varphi_{X_i}(t/n)]^n \quad (\text{since } X_i \text{'s are i.i.d.}) \\ &= [e^{-|t/n|}]^n = e^{-|t|} = \varphi_X(t) \end{aligned}$$

Hence by uniqueness theorem of characteristic functions, we have:

"The arithmetic mean  $\bar{X}$  of a sample  $X_1, X_2, \dots, X_n$  of independent observations from a standard Cauchy distribution is also a standard Cauchy variate. In other words, the arithmetic mean of a random sample of any size yields exactly as much information as a single determination of  $X$ ."

This implies that the sample mean  $\bar{X}_n$  of a random sample of size  $n$ , as an estimate of population mean does not improve with increasing  $n$ , which contradicts the Weak Law of Large Numbers (WLLN).

### 8.9.2. Moments of Cauchy Distribution.

$$E(Y) = \int_{-\infty}^{\infty} yf(y) dy = \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{y}{\lambda^2 + (y - \mu)^2} dy$$

$$\begin{aligned}
 &= \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{(y - \mu) + \mu}{\lambda^2 + (y - \mu)^2} dy \\
 &= \mu \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{dy}{\lambda^2 + (y - \mu)^2} + \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{(y - \mu)}{\lambda^2 + (y - \mu)^2} dy \\
 &= \mu \cdot 1 + \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{z}{\lambda^2 + z^2} dz
 \end{aligned}$$

Although the integral  $\int_{-\infty}^{\infty} \frac{z}{\lambda^2 + z^2} dz$ , is not completely convergent, i.e.,

$$\lim_{n \rightarrow \infty} \int_{-n}^{n'} \frac{z}{\lambda^2 + z^2} dz \text{ does not exist, its principal value, viz., } \lim_{n' \rightarrow \infty - n} \int_{-n}^n \frac{z}{\lambda^2 + z^2} dz$$

exists and is equal to zero. Thus, in the general sense the mean of Cauchy distribution does not exist. But, if we conventionally agree to assume that the mean of Cauchy distribution exists (by taking the principal value), then it is located at  $x = \mu$ . Also, obviously, the probability curve is symmetrical about the point  $x = \mu$ . Hence for this distribution, the mean, median and mode coincide at the point  $x = \mu$ .

$$\mu_2 = E(Y - \mu)^2 = \int_{-\infty}^{\infty} (y - \mu)^2 f(y) dy = \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{(y - \mu)^2}{\lambda^2 + (y - \mu)^2} dy,$$

which does not exist since the integral is not convergent. Thus, in general, for the Cauchy's distribution  $\mu_r$ , ( $r \geq 2$ ) do not exist.

**Remark.** The role of Cauchy distribution in statistical theory often lies in providing counter examples, e.g. it, is often quoted as a distribution for which moments do not exist. It also provides an example to show that

$$\varphi_{X+Y}(t) = \varphi_X(t) \varphi_Y(t)$$

does not imply that  $X$  and  $Y$  are independent. [See Remark to Theorem 6.23]

Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a standard Cauchy distribution. Let  $\bar{X} = \sum_{i=1}^n X_i / n$ . Since  $E(X_i)$  does not exist ( $\because$  mean of a Cauchy

distribution does not exist),  $E(\bar{X})$  does not exist either and the definition of an unbiased estimate does not apply to  $\bar{X}$ .

Cauchy distribution also contradicts the WLLN [See Remark 4, § 8.9.1].

**Example 8.42.** Let  $X$  have a (standard) Cauchy distribution. Find a p.d.f. for  $X^2$  and identify its distribution. [Delhi Univ. B.Sc. (Stat. Hons.), 1989; '87]

**Solution.** Since  $X$  has a standard Cauchy distribution, its p.d.f. is

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, -\infty < x < \infty$$

The distribution function of  $Y = X^2$  is

$$\begin{aligned} G_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} f(x) dx = 2 \frac{1}{\pi} \int_0^{\sqrt{y}} \frac{dx}{1+x^2} \\ &= \frac{2}{\pi} \tan^{-1}(\sqrt{y}), \quad 0 < y < \infty \end{aligned}$$

The p.d.f.  $g_Y(y)$  of  $Y$  is given by

$$\begin{aligned} g_Y(y) &= \frac{d}{dy} [G_Y(y)] = \frac{2}{\pi} \cdot \frac{1}{(1+y)} \cdot \frac{1}{2\sqrt{y}} \\ &= \frac{1}{\pi} \cdot \frac{y^{-1/2}}{1+y} = \frac{1}{B(\frac{1}{2}, \frac{1}{2})} \cdot \frac{y^{(1/2)-1}}{(1+y)^{(1/2+1/2)}}, \quad y > 0 \end{aligned}$$

This is the p.d.f. of Beta distribution of second kind with parameters  $(\frac{1}{2}, \frac{1}{2})$ , i.e.,  $X^2 \sim \beta_2(\frac{1}{2}, \frac{1}{2})$

Remark. Here  $y = g(x) = x^2$ , gives  $g'(x) = 2x$  which is sometimes  $> 0$  and sometimes  $< 0$ . Hence Theorem 5.9 can not be used in this case.

Example 8.43. Let  $X \sim N(0, 1)$  and  $Y \sim N(0, 1)$  be independent random variables. Find the distribution of  $X/Y$  and identify it.

[Delhi Univ. B.Sc. (Stat. Hons.), 1990; Nagpur Univ. B.Sc., 1991]

Solution. Since  $X$  and  $Y$  are independent  $N(0, 1)$ , their joint p.d.f. is given by

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y) = \frac{1}{2\pi} \cdot e^{-(x^2+y^2)/2}$$

Let us make the following transformation of variables

$$u = x/y, v = y \text{ so that } x = uv, y = v$$

Jacobian of transformation  $J = v$ .

Hence the joint p.d.f. of  $U$  and  $V$  becomes

$$\begin{aligned} g_{UV}(u, v) &= \frac{1}{2\pi} \cdot \exp \left\{ -(u^2 v^2 + v^2)/2 \right\} |J| \\ &= \frac{1}{2\pi} \exp \left\{ -(1+u^2) v^2/2 \right\} |v|, \quad -\infty < (u, v) < \infty \end{aligned}$$

The marginal p.d.f. of  $U$  is

$$\begin{aligned} g_U(u) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp \left\{ -(1+u^2) v^2/2 \right\} |v| dv \\ &= \frac{1}{\pi} \int_0^{\infty} e^{-t} \frac{dt}{(1+u^2)} \quad \left[ \left( \frac{1}{2}(1+u^2) v^2 = t \right) \right] \\ &= \frac{1}{\pi(1+u^2)} \left| -e^{-t} \right|_0^{\infty} = \frac{1}{\pi(1+u^2)}, \quad -\infty < u < \infty \end{aligned}$$

which is the p.d.f. of a standard Cauchy distribution.

Thus the ratio of two independent standard normal variates is a standard Cauchy variate.

**Example 8.44.** Let  $X$  and  $Y$  be i.i.d. standard Cauchy variates. Prove that the p.d.f. of  $XY$  is :  $\frac{2}{\pi^2} \left\{ \frac{\log|x|}{x^2 - 1} \right\}$ . [Delhi Univ. M.Sc. (Stat), 1991]

**Solution.** Since  $X$  and  $Y$  are independent standard Cauchy variates, their joint p.d.f. is given by

$$f(x, y) = \frac{1}{\pi^2} \cdot \frac{1}{(1+x^2)(1+y^2)} ; -\infty < (x, y) < \infty .$$

Let  $u = xy$  and  $v = y$ . Then Jacobian of transformation is given by

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{1}{v} & -\frac{u}{v^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{v} \quad \left( \because y = v, x = \frac{u}{v} \right)$$

Thus the joint p.d.f. of  $U$  and  $V$  is given by

$$\begin{aligned} g(u, v) &= \frac{1}{\pi^2} \frac{1}{\left(1 + \frac{u^2}{v^2}\right)(1+v^2)} \cdot \frac{1}{|v|} \\ &= \frac{1}{\pi^2} \frac{|v|}{(u^2 + v^2)(1+v^2)} ; -\infty < (u, v) < \infty \end{aligned}$$

Integrating w.r.to  $v$  over the range  $-\infty$  to  $\infty$ , the marginal p.d.f. of  $U$  is given by

$$\begin{aligned} g_1(u) &= \int_{-\infty}^{\infty} g(u, v) dv = \frac{1}{\pi^2} \int_{-\infty}^{\infty} \frac{|v|}{(u^2 + v^2)(1+v^2)} dv \\ &= \frac{2}{\pi^2} \int_0^{\infty} \frac{|v|}{(u^2 + v^2)(1+v^2)} dv , \end{aligned}$$

(Since the integrand is an even function of  $v$ .)

$$\begin{aligned} \therefore g_1(u) &= \frac{2}{\pi^2} \int_0^{\infty} \frac{v}{(u^2 + v^2)(1+v^2)} dv \\ &= \frac{1}{\pi^2} \int_0^{\infty} \frac{2v}{(u^2 - 1)} \left( \frac{1}{1+v^2} - \frac{1}{u^2+v^2} \right) dv \end{aligned}$$

$$\begin{aligned} &= \frac{1}{\pi^2 (u^2 - 1)} \left| \log(1+v^2) - \log(u^2+v^2) \right|_0^{\infty} \\ &= \frac{1}{\pi^2 (u^2 - 1)} \left| \log \left( \frac{1+v^2}{u^2+v^2} \right) \right|_0^{\infty} \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{\pi^2(u^2 - 1)} \left\{ \left[ \log \left( \frac{\frac{1}{v^2} + 1}{\frac{u^2}{v^2} + 1} \right) \right]_{v \rightarrow \infty} - \log \left( \frac{1}{u^2} \right) \right\} \\
 &= \frac{1}{\pi^2(u^2 - 1)} [\log 1 + 2 \log |u|] = \frac{2 \log |u|}{\pi^2(u^2 - 1)}, -\infty < u < \infty
 \end{aligned}$$

### EXERCISE 8(f)

1. (a) Show that a function

$$f(x; \mu, \lambda) = \frac{k}{\lambda^2 + (x - \mu)^2}; -\infty < x < \infty$$

represents a frequency function of a distribution for a suitable value of  $k$ . Determine  $k$  and obtain median and quartiles of the distribution. Hence interpret the parameters  $\lambda$  and  $\mu$  of the distribution.

(b) If  $X$  is a Cauchy variate with parameters  $\lambda$  and  $\mu$ , find the characteristic function  $\varphi_X(t)$ . Discuss briefly the role of Cauchy distribution in Statistics.

[Bombay Univ. B.Sc. (Stat.), 1993]

(c) "The role of Cauchy distribution often lies in providing counter examples." Justify. [Delhi Univ. B.Sc. (Stat. Hons.), 1991, '88]

(d) Discuss briefly the role of Cauchy distribution in statistics:

If  $X_1, X_2, \dots, X_n$  are independent standard Cauchy variates, show that the mean  $\bar{X} = (X_1 + X_2 + \dots + X_n)/n$ , is also a Cauchy variate.

[Delhi Univ. B.Sc. (Stat. Hons.), 1986]

2. (a) If  $X$  and  $Y$  are independent random variables following Cauchy distribution with parameters  $(\lambda_1, \mu_1)$  and  $(\lambda_2, \mu_2)$  respectively, show that  $X + Y$  follows Cauchy distribution with parameters  $\lambda_1 + \lambda_2$  and  $\mu_1 + \mu_2$ .

(b) Obtain the characteristic function of Cauchy distribution

$$dF(x) = \frac{dx}{\pi(1+x^2)}, -\infty < x < \infty$$

If  $X_1, X_2, \dots, X_n$  are independent Cauchy variates, show that the mean  $\bar{X} = \frac{1}{n} \sum X$  is also a Cauchy variate.

3. Let  $X$  and  $Y$  be standard normal variates. Find the distribution of  $U = X/|Y|$ .

$$\text{Ans. } f(u) = \frac{1}{\pi} \cdot \frac{1}{1+u^2}, -\infty < u < \infty$$

4. A needle spins about the point  $(0, b)$  of the  $x$ - $y$  plane with  $b > 0$  and comes to a stop thereby making an angle  $\varphi$  with  $Y$ -axis. The direction of the needle then intersects the  $x$ -axis at a point  $(X, 0)$ . Assuming  $\varphi$  is a r.v. with uniform

probability distribution on  $(-\pi/2, \pi/2)$ , what is the distribution function and hence p.d.f. of  $X$ ?

$$\text{Ans. } F_X(x) = \frac{1}{\pi} \left[ \tan^{-1}(x/b) + \frac{\pi}{2} \right], f_X(x) = \frac{1}{\pi} \cdot \frac{b}{x^2 + b^2}, -\infty < X < \infty$$

5. If  $X_1, X_2, X_3, X_4$  are independent standard normal variates, find the distribution of  $\frac{X_1}{X_2} + \frac{X_3}{X_4}$ .

6.  $\bar{X}_n$  is the mean of  $n$  independent random variables distributed like  $X$ , and  $X$  has a symmetric distribution. If  $\bar{X}_n$  has exactly the same distribution as  $X$  for all  $n$ , then prove that the characteristic function of  $X$  is

$$\Phi_X(t) = e^{-c|t|}$$

for some real constant  $c > 0$ . Identify this distribution.

7. If  $X \sim N(\mu_1, \sigma_1^2)$  and  $Y \sim N(\mu_2, \sigma_2^2)$  are independent random variables, obtain the p.d.f. of  $U = \frac{X - \mu_1}{Y - \mu_2}$ . (I.I.T., B.Tech. 1993)

**8-10. Central Limit Theorem.** The central limit theorem in the mathematical theory of probability may be expressed as follows :

"If  $X_i$ , ( $i = 1, 2, \dots, n$ ), be independent random variables such that  $E(X_i) = \mu_i$  and  $V(X_i) = \sigma_i^2$ , then it can be proved that under certain very general conditions, the random variable  $S_n = X_1 + X_2 + \dots + X_n$ , is asymptotically normal with mean  $\mu$  and standard deviation  $\sigma$  where

$$\mu = \sum_{i=1}^n \mu_i \quad \text{and} \quad \sigma^2 = \sum_{i=1}^n \sigma_i^2$$

This theorem was first stated by Laplace in 1812 and a rigorous proof under fairly general conditions was given by Liapounoff in 1901. Below we shall consider some particular cases of this general central limit theorem.

**De-Moivre's-Laplace theorem. (1733).** A particular case of central limit theorem is De-Moivre's theorem which states as follows:

"If  $X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } q \end{cases}$

then the distribution of the random variable  $S_n = X_1 + X_2 + \dots + X_n$ , where  $X_i$ 's are independent, is asymptotically normal as  $n \rightarrow \infty$ ."

**Proof.** M.G.F. of  $X_i$  is given by

$$M_{X_i}(t) = E(e^{tX_i}) = e^{t \cdot 1} p + e^{t \cdot 0} q = (q + pe^t)$$

M.G.F. of the sum  $S_n = X_1 + X_2 + \dots + X_n$  is given by

$$M_{S_n}(t) = M_{X_1 + X_2 + \dots + X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \dots M_{X_n}(t)$$

$$= [M_{X_i}(t)]^n \quad (\text{since } X_i \text{'s are identically distributed})$$

$$= (q + pe^t)^n,$$

which is the M.G.F. of a binomial variate with parameters  $n$  and  $p$ .

$$\therefore E(S_n) = np = \mu \text{ (say), and } V(S_n) = npq = \sigma^2, \text{ (say).}$$

Let

$$Z = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - \mu}{\sigma}$$

$$M_Z(t) = e^{-\mu t/\sigma} M_{S_n}(t/\sigma) \quad [\text{c.f. Chapter 6}]$$

$$= e^{-\mu t/\sqrt{npq}} \left[ q + p e^{t/\sqrt{npq}} \right]^n$$

$$= \left[ 1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n \quad [\text{c.f. Example 7-19}]$$

where  $O(n^{-3/2})$  represents terms involving  $n^{3/2}$  and higher powers of  $n$  in the denominator.

Proceeding to the limits as  $n \rightarrow \infty$ , we get

$$\lim_{n \rightarrow \infty} M_Z(t) = \lim_{n \rightarrow \infty} \left[ 1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n = \lim_{n \rightarrow \infty} \left[ 1 + \frac{t^2}{2n} \right]^n = e^{t^2/2}$$

which is the M.G.F. of a standard normal variate. Hence by the uniqueness theorem of M.G.F.'s

$$Z = \frac{S_n - \mu}{\sigma} \text{ is asymptotically } N(0, 1).$$

Hence  $S_n = X_1 + X_2 + \dots + X_n$  is asymptotically  $N(\mu, \sigma^2)$  as  $n \rightarrow \infty$ .

**Remarks 1.** From this theorem it follows that standard binomial variate tends to standard normal variate as  $n \rightarrow \infty$ . In other words, binomial distribution tends to normal distribution as  $n \rightarrow \infty$ .

**2. Convergence in Distribution or Law.** Let  $\{X_n\}$  be a sequence of r.v.'s and  $\{F_n\}$  be the corresponding sequence of distribution functions. We say that  $X_n$  converges in distribution (or law) to  $X$  if there exists a r.v.  $X$  with distribution function  $F$  such that as  $n \rightarrow \infty$ ,  $F_n(x) \rightarrow F(x)$  at every point  $x$  at which  $F$  is continuous.

We write  $X_n \xrightarrow{L} X$  or  $X_n \xrightarrow{d} X$ .

**3.** It may be remarked that convergence in probability discussed in § 6-14 implies convergence in distribution (or law) i.e.,

$$X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L} X \quad ...(*)$$

The converse is not true i.e.,  $X_n \xrightarrow{L} X$ , in general, does not imply  $X_n \xrightarrow{P} X$

However, we have the following result.

Let  $k$  be a constant. Then

$$X_n \xrightarrow{L} k \Rightarrow X_n \xrightarrow{P} k \quad ...(**)$$

Combining (\*) and (\*\*), we get the following result.

Let  $k$  be a constant. Then

$$X_n \xrightarrow{L} k \Leftrightarrow X_n \xrightarrow{P} k \quad ...(***)$$

**8-10-1 Lindeberg-Levy Theorem.** The following case of central limit theorem for equal components, i.e., for identically distributed variables, was first proved by Lindeberg and Levy.

"If  $X_1, X_2, \dots, X_n$  are independently and identically distributed random variables with

$$\left. \begin{array}{l} E(X_i) = \mu_1 \\ V(X_i) = \sigma_1^2 \end{array} \right\} i = 1, 2, \dots, n$$

then the sum  $S_n = X_1 + X_2 + \dots + X_n$  is asymptotically normal with mean  $\mu = n\mu_1$  and variance  $\sigma^2 = n\sigma_1^2$ .

Here we make the following assumptions :

(i) The variables are independent and identically distributed

(ii)  $E(X_i^2)$  exists for all  $i = 1, 2, \dots$

**Proof.** Let  $M_1(t)$  denote the M.G.F. of each of the deviation  $(X_i - \mu_1)$  and  $M(t)$  denote the M.G.F. of the standard variate

$$Z = (S_n - \mu)/\sigma$$

Since  $\mu_1'$  and  $\mu_2'$ , (about origin) of the deviation  $(X_i - \mu_1)$  are given by

$$\mu_1' = E(X_i - \mu_1) = 0, \mu_2' = E(X_i - \mu_1)^2 = \sigma_1^2$$

We have

$$\begin{aligned} M_1(t) &= \left( 1 + \mu_1't + \mu_2'\frac{t^2}{2!} + \mu_3'\frac{t^3}{3!} + \dots \right) \\ &= \left[ 1 + \frac{t^2}{2!} \sigma_1^2 + O(t^3) \right] \end{aligned} \quad \dots (*)$$

where  $O(t^3)$  contains terms with  $t^3$  and higher powers of  $t$ .

We have

$$Z = \frac{S_n - \mu}{\sigma} = \frac{(X_1 + X_2 + \dots + X_n) - n\mu_1}{\sigma} = \sum_{i=1}^n \left( \frac{X_i - \mu_1}{\sigma} \right)$$

and since  $X_i$ 's are independent, we get

$$\begin{aligned} M_Z(t) &= M \sum_{i=1}^n (X_i - \mu_1)/\sigma_1 (t) = M \sum_{i=1}^n (X_i - \mu_1) (t/\sigma) \\ &= \prod_{i=1}^n \left[ M_{(X_i - \mu_1)}(t/\sigma) \right] = [M_1(t/\sigma)]^n = [M_1(t/\sqrt{n}\sigma_1)]^n \\ &= \left[ 1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n \quad [\text{From } (*)] \end{aligned}$$

For every fixed 't', the terms  $O(n^{-3/2}) \rightarrow 0$  as  $n \rightarrow \infty$ . Therefore, as  $n \rightarrow \infty$ , we get

$$\lim_{n \rightarrow \infty} M_Z(t) = \lim_{n \rightarrow \infty} \left[ 1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n = \exp [t^2/2],$$

which is the M.G.F. of standard normal variate. Hence by uniqueness theorem of M.G.F.'s  $Z = (S_n - \mu)/\sigma$  is asymptotically  $N(0, 1)$ , or  $S_n = X_1 + X_2 + \dots + X_n$  is asymptotically  $N(\mu, \sigma^2)$ , where  $\mu = n\mu_1$  and  $\sigma^2 = n\sigma_1^2$ .

Note. C.L.T. can be stated in another form as follows:

(i) If  $X_1, X_2, \dots, X_n$  are i.i.d. with mean  $\mu_1$  and variance  $\sigma_1^2$  (finite) and  $S_n = X_1 + X_2 + \dots + X_n$ , then

$$\lim_{n \rightarrow \infty} P \left[ a \leq \frac{S_n - n\mu_1}{\sigma_1 \sqrt{n}} \leq b \right] = \Phi(b) - \Phi(a) \\ = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \quad \dots(8.31)$$

for  $-\infty < a < b < \infty$ ;  $\Phi(-\infty) = 0$ ,  $\Phi(\infty) = 1$

or

$$(ii) \lim_{n \rightarrow \infty} P \left[ a \leq \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \leq b \right] = \Phi(b) - \Phi(a) \quad \dots(8.31.1)$$

or, still another form :

$$(iii) \lim_{n \rightarrow \infty} P \left[ a \leq \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var(\bar{X}_n)}} \leq b \right] = \Phi(b) - \Phi(a)$$

$$i.e., \lim_{n \rightarrow \infty} P \left[ a \leq \frac{\bar{X}_n - \mu_1}{\sigma_1 / \sqrt{n}} \leq b \right] = \Phi(b) - \Phi(a) \quad \dots(8.31.2)$$

**Remarks 1.** We wrote the C.L.T. using non-strict inequalities

$$P[a \leq (\cdot) \leq b]$$

It makes no difference whether one or both are changed to a strict inequality. The reason is that the limit distribution function (d.f.)  $\Phi(\cdot)$  is a continuous d.f.

2. In the binomial case, C.L.T. gives good approximation if  $p$  is nearly 1/2. For  $p$  near about 0 or 1, the C.L.T. approximation still holds but in that case  $n$  has to be sufficiently greater than in the case  $p = 1/2$  approximately.

**8.10.2. Applications of Central Limit Theorem.** (a) If  $X_1, X_2, \dots$  are i.i.d.  $B(r, p)$  and

$$S_n = X_1 + X_2 + \dots + X_n, \text{ then}$$

$$E(S_n) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n (rp) = nrp$$

$$\text{and} \quad V(S_n) = V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i) = \sum_{i=1}^n (rpq) = nrpq$$

Hence (8.31.1)

$$\Rightarrow \lim_{n \rightarrow \infty} P \left[ a \leq \frac{S_n - nrp}{\sqrt{nrp(1-p)}} \leq b \right] = \Phi(b) - \Phi(a), 0 < p < 1$$

(b) If  $Y_n$  is binomial variate with parameters  $n$  and  $p$ , then

$$\lim_{n \rightarrow \infty} P \left[ a \leq \frac{Y_n - np}{\sqrt{np(1-p)}} \leq b \right] = \Phi(b) - \Phi(a), \quad 0 < p < 1$$

**Proof.** Let  $X_1, X_2, \dots$  be i.i.d. Bernoulli variates, i.e.,  $B(1, p)$ , then  $S_n = X_1 + X_2 + \dots + X_n = B(n, p)$ . But  $Y_n = B(n, p)$

Hence using  $Y_n$  instead of  $S_n$  in (8-31-1), we get

$$\lim_{n \rightarrow \infty} P \left[ a \leq \frac{Y_n - E(Y_n)}{\sqrt{\text{Var } Y_n}} \leq b \right] = \Phi(b) - \Phi(a)$$

$$\text{i.e.,} \quad \lim_{n \rightarrow \infty} P \left[ a \leq \frac{Y_n - np}{\sqrt{npq}} \leq b \right] = \Phi(b) - \Phi(a), \quad q = 1 - p$$

(c) If  $Y_n$  is distributed as  $P(n)$ , then

$$\lim_{n \rightarrow \infty} P \left[ a \leq \frac{Y_n - n}{\sqrt{n}} \leq b \right] = \Phi(b) - \Phi(a)$$

Thus, for instance

$$\lim_{n \rightarrow \infty} P(Y_n \leq n) = \frac{1}{2}, \text{ i.e., } \sum_{k=0}^n \frac{e^{-n} n^k}{k!} = \frac{1}{2} \text{ as } n \rightarrow \infty$$

**Proof.** Let  $X_1, X_2, \dots$  be i.i.d.  $P(1)$ . Then

$$S_n = X_1 + X_2 + \dots + X_n \sim P(n) \Rightarrow Y_n = S_n$$

$$\therefore P \left[ a \leq \frac{Y_n - n}{\sqrt{n}} \leq b \right] = P \left[ a \leq \frac{S_n - n}{\sqrt{n}} \leq b \right] \rightarrow \Phi(b) - \Phi(a) \text{ as } n \rightarrow \infty$$

In particular, let us take  $a = -\infty$  and  $b = 0$ , then

$$P \left( a \leq \frac{Y_n - n}{\sqrt{n}} \leq b \right) = P \left( \frac{Y_n - n}{\sqrt{n}} \leq 0 \right) = P(Y_n \leq n) \quad \dots (*)$$

$$\text{Also } \Phi(b) - \Phi(a) = \Phi(0) - \Phi(-\infty) = 1/2 \quad \dots (**)$$

From (\*) and (\*\*), we get

$$P(Y_n \leq n) \rightarrow 1/2 \text{ as } n \rightarrow \infty$$

**Remark.** This result could be generalised. In fact, on taking  $a = -\infty$  and  $b = 0$  in (8-31-1), we get

$$P \left[ \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} \leq 0 \right] \rightarrow 1/2 \Rightarrow P[S_n \leq E(S_n)] \rightarrow 1/2 \text{ as } n \rightarrow \infty$$

**8-10-3. Liapounoff's Central Limit Theorem.** Below we shall give (without proof) the central limit theorem for the generalised case when the variables are not identically distributed and where, in addition to the existence of the second moment for the variables  $X_i$ , we impose some further conditions.

Let  $X_1, X_2, \dots, X_n$  be independent random variables such that

$$\begin{aligned} E(X_i) &= \mu_i \\ V(X_i) &= \sigma_i^2 \end{aligned} ; \quad i = 1, 2, \dots, n.$$

Let us suppose that third absolute moment of  $X_i$  about its mean viz.,

$$\rho_i^3 = E \{ |X_i - \mu_i|^3 \}; i = 1, 2, \dots, n$$

is finite. Further let

$$\rho^3 = \sum_{i=1}^n \rho_i^3$$

If  $\lim_{n \rightarrow \infty} \frac{\rho}{\sigma} = 0$ , the sum  $X = X_1 + X_2 + \dots + X_n$  is asymptotically

$$N(\mu, \sigma^2), \text{ where } \mu = \sum_{i=1}^n \mu_i \text{ and } \sigma^2 = \sum_{i=1}^n \sigma_i^2.$$

**Remarks.** 1. (About Liapounoff's theorem). If the variables  $X_i; i = 1, 2, \dots, n$  are identical, then

$$\rho^3 = \sum_{i=1}^n \rho_i^3 = n \rho_1^3 \text{ and } \sigma^2 = \sum_{i=1}^n \sigma_i^2 = n \sigma_1^2$$

$$\therefore \frac{\rho}{\sigma} = \frac{n^{1/3} \rho_1}{n^{1/2} \cdot \sigma_1} = \frac{\rho_1}{\sigma_1} \cdot \frac{1}{n^{1/6}} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Thus for identical variables, the condition of Liapounoff's theorem is satisfied.

It may be pointed out here that Lindeberg - Levy theorem proved in § 8-10-1, should not be inferred as a particular case of Liapounoff's theorem, since the former does not assume the existence of the third moment.

2. Central limit theorem can be expected in the following cases:

(i) If a certain random variable  $X$  arises as cumulative effect of several independent causes each of which can be considered as a continuous random variable, then  $X$  obeys central limit theorem under certain regularity conditions.

(ii) If  $\varphi(X_1, X_2, \dots, X_n)$ , is a function of  $X_i$ 's having first and second continuous derivatives about the point  $(\mu_1, \mu_2, \dots, \mu_n)$ , then under certain regularity conditions,  $\varphi(X_1, X_2, \dots, X_n)$  is asymptotically normal with mean  $\varphi(\mu_1, \mu_2, \dots, \mu_n)$ .

(iii) Under certain conditions, the central limit theorem holds for variables which are not independent.

3. Relation between CLT and WLLN. (a). Both the central limit theorem (CLT) and the weak law of large numbers hold for a sequence  $\{X_n\}$  of i.i.d. random variables with finite mean  $\mu$  and variance  $\sigma^2$ .

However, in this case the CLT is a stronger result than the WLLN in the sense that the former provides an estimate of the  $P[|S_n - n\mu| / n \geq \varepsilon]$ , as given below:

$$\begin{aligned} P \left[ \left| \frac{S_n - n\mu}{n} \right| \geq \varepsilon \right] &= P \left[ \left| \bar{X}_n - \mu \right| \geq \varepsilon \right] \\ &= P \left[ \left| \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right| \geq \frac{\varepsilon}{\sigma/\sqrt{n}} \right] \\ &= P[|Z| \geq \varepsilon \sqrt{n}/\sigma]; Z \sim N(0, 1) \end{aligned}$$

$$\begin{aligned} &= 1 - P[|Z| \leq \epsilon \sqrt{n}/\sigma] \\ &= 1 - \left[ \Phi\left(\frac{\epsilon \sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\epsilon \sqrt{n}}{\sigma}\right) \right] \end{aligned}$$

where  $\Phi(\cdot)$  is the distribution function of standard normal variate.

However, WLLN does not require the existence of variance (c.f. Khinchin's theorem).

(b) For the sequence  $\{X_n\}$  of independent and uniformly bounded r.v.'s, WLLN holds [c.f. theorem 6-32] and CLT holds in this case provided

$$B_n = \text{Var}(X_1 + X_2 + \dots + X_n) = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2 \rightarrow \infty \text{ as } n \rightarrow \infty.$$

(c) For the sequence  $\{X_n\}$  of independent r.v.'s, CLT may hold but the WLLN may not hold.

**8-10-4. Cramér's Theorem.** We state below (without proof), a useful result on the convergence of sequences of r.v.'s.

**Cramér's Theorem.** Let  $\{X_n\}$  and  $\{Y_n\}$  be sequences of r.v.'s such that:

$$X_n \xrightarrow{L} X \text{ and } Y_n \xrightarrow{P} c \text{ (constant),}$$

$$\text{then } \frac{X_n}{Y_n} \xrightarrow{L} \frac{X}{c} \text{ if } c \neq 0$$

For illustrations, see Example 8-46 and Qns. 15 to 17 in Exercise 8 (g).

**Example 8-45.** Let  $X_1, X_2, \dots$  be a i.i.d. Poisson variates with parameter  $\lambda$ . Use CLT to estimate  $P(120 \leq S_n \leq 160)$ , where

$$S_n = X_1 + X_2 + \dots + X_n; \lambda = 2 \text{ and } n = 75.$$

**Solution.** Since  $X_i$  is i.i.d.  $P(\lambda)$ ,

$$E(X_i) = \lambda \text{ and } \text{Var}(X_i) = \lambda; i = 1, 2, \dots, n$$

$$\therefore E(S_n) = \sum_{i=1}^n E(X_i) = n\lambda$$

$$\text{Var}(S_n) = \text{Var}(X_1 + X_2 + \dots + X_n) = \sum_{i=1}^n \text{Var} X_i = n\lambda$$

Hence by Lindeberg-Levy CLT, (for large  $n$ )

$$S_n \sim N(n\lambda, n\lambda) = N(\mu = 150, \sigma^2 = 150); (n = 75; \lambda = 2)$$

$$\begin{aligned} \therefore P(120 \leq S_n \leq 160) &= P\left(\frac{120 - 150}{\sqrt{150}} \leq Z \leq \frac{160 - 150}{\sqrt{150}}\right) \\ &= P(-2.45 \leq Z \leq 0.82); Z \sim N(0, 1) \\ &= P(-2.45 \leq Z \leq 0) + P(0 \leq Z \leq 0.82) \\ &= P(0 \leq Z \leq 2.45) + P(0 \leq Z \leq 0.82) \end{aligned}$$

**Example 8-46.** Let  $X_1, X_2, \dots, X_n$  be i.i.d. standardised variates with  $E(X_i^4) < \infty$ . Find the limiting distribution of:

$$Z_n = \sqrt{n} [X_1 X_2 + X_3 X_4 + \dots + X_{2n-1} X_{2n}] \div [X_1^2 + X_2^2 + \dots + X_{2n}^2]$$

**Solution.** Since  $X_i$ 's are i.i.d. standardised variates we have:

$$E(X_i) = 0; \quad \text{Var}(X_i) = E(X_i^2) = 1, \quad i = 1, 2, \dots, n \quad \dots (*)$$

Let  $Y_i = X_{2i-1} X_{2i}; \quad i = 1, 2, \dots, n$

$$\Rightarrow E(Y_i) = E(X_{2i-1}) E(X_{2i}) = 0 \quad (\because X_i \text{'s are independent})$$

$$\therefore \text{Var}(Y_i) = E(Y_i^2) = E(X_{2i-1}^2 X_{2i}^2) = E(X_{2i-1}^2) E(X_{2i}^2) = 1$$

Hence  $Y_i, i = 1, 2, \dots, n$  are also i.i.d. standardised variates. Hence by CLT

for i.i.d. r.v.'s,  $\left[ S_n = \sum_{i=1}^n Y_i \right]$ , we get

$$U_n = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{X_1 X_2 + X_3 X_4 + \dots + X_{2n-1} X_{2n}}{\sqrt{n}} \xrightarrow{L} Z \sim N(0, 1)$$

as  $n \rightarrow \infty$

$$\text{Also } E(X_i^2) = 1 \text{ (finite), } i = 1, 2, \dots, n.$$

Hence by Khinchine's theorem, WLLN applies to the sequence

$\{X_i^2\}, i = 1, 2, \dots, 2n$ ; so that

$$V_n = \frac{X_1^2 + X_2^2 + \dots + X_{2n}^2}{2n} \xrightarrow{P} E(X_i^2) = 1, \text{ as } n \rightarrow \infty.$$

Hence by Cramer's theorem

$$\lim_{n \rightarrow \infty} \frac{U_n}{V_n} = \frac{2\sqrt{n} [X_1 X_2 + X_3 X_4 + \dots + X_{2n-1} X_{2n}]}{X_1^2 + X_2^2 + \dots + X_{2n}^2} \xrightarrow{L} \frac{Z}{1} \sim N(0, 1)$$

$$\Rightarrow \lim_{n \rightarrow \infty} \frac{\sqrt{n} [X_1 X_2 + X_3 X_4 + \dots + X_{2n-1} X_{2n}]}{X_1^2 + X_2^2 + \dots + X_{2n}^2} \xrightarrow{L} \frac{Z}{2} \sim N(0, 1/4)$$

$$[\because Z \sim N(0, 1) \Rightarrow CZ \sim N(0, C^2)]$$

### EXERCISE 8(g)

1. State and prove the central limit theorem for the sum of  $n$  independently and identically distributed random variables with positive finite variance under conditions to be stated.

2. State Lindeberg's sufficient conditions for the central limit theorem to hold for a sequence  $\{X_k\}$  of independent random variables. Show that every uniformly bounded sequence  $\{X_k\}$  of mutually independent random variables obeys the central limit theorem.

Comment on the case when the random variables do not possess expectations.

3. A distribution with unknown mean  $\mu$  has variance equal to 1.5. Use central limit theorem to find how large a sample should be taken from the

distribution in order that the probability will be at least 0.95 that the sample mean will be within 0.5 of the population mean.

4. The life time of a certain brand of an electric bulb may be considered a random variable with mean 1,200 hours and standard deviation 250 hours. Find the probability, using central limit theorem, that the average life-time of 60 bulbs exceeds 1,400 hours.

5. State the Liapounoff form of central limit theorem.

Decide whether the central limit theorem holds for the sequence of independent random variables  $X_r$  with distribution defined as follows:

$$P(X_r = 1) = p_r \text{ and } P(X_r = 0) = 1 - p_r$$

6. Show that the central limit theorem applies if

$$(i) P(X_k = \pm k^\alpha) = \frac{1}{2}, \quad (ii) P(X_k = \pm \sqrt{2k-1}) = \frac{1}{2}, \text{ and}$$

$$(iii) P(X_k = 0) = 1 - k^{1-2\alpha}, P(X_k = \pm k^\alpha) = \frac{1}{2} k^{-2\alpha}, \text{ where } \alpha < \frac{1}{2}$$

7. If  $X_1, X_2, X_3, \dots$  is a sequence of independent random variables having the uniform densities

$$f_i(x_i) = \begin{cases} 1/(2-i^{-1}), & 0 < x_i < 2 - i^{-1} \\ 0 & \text{elsewhere,} \end{cases}$$

show that the central limit theorem holds.

8. Let  $X_n$  be the sample mean of a random sample of size  $n$  from Rectangular distribution on  $[0, 1]$ . Let

$$U_n = \sqrt{n} (\bar{X}_n - \frac{1}{2}).$$

$$\text{Show that } F(u) = \lim_{n \rightarrow \infty} P(U_n < u)$$

exists and determine it.

$$\text{Ans. } \Phi(\sqrt{12} u), \text{ where } \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-x^2/2} dx$$

9. Let  $X_1, X_2, \dots$  be a sequence of independent, identically distributed non-negative random variables such that  $E(\log X_1)^2$  is finite.  $Z_n = (X_1 X_2 \dots X_n)^{1/n}$ . Show that the positive constant  $c$  can be so chosen that the random variable  $(cZ_n)^{\sqrt{n}}$  has a non-degenerate limit distribution function  $F(\cdot)$  and determine  $F(\cdot)$ .

$$\text{Ans. } c = e^{-\mu}, F(x) = \Phi(\log x/\sigma), \mu = E(\log X_1) \text{ and } \sigma^2 = V(\log X_1).$$

10.  $\{X_n\}$  is a sequence of *i.i.d.* random variables. If  $n$  is a perfect square, then  $X_n$  is a Cauchy variate with density  $\frac{1}{\pi} \cdot \frac{1}{1+x^2}, -\infty < x < \infty$ .

Otherwise  $X_n$  has a distribution function  $F(x)$  with mean zero and finite variance  $\sigma^2$ . Discuss the asymptotic distribution of  $(X_1 + X_2 + \dots + X_n)/\sqrt{n}$ .

11. Let  $\{X_k\}, k \geq 1$  be a sequence of *i.i.d.* variates with

$$f(x) = \frac{1}{2} e^{-|x|}, -\infty < x < \infty.$$

Find the constants  $a_n$  and  $b_n$  such that

$$\{ |X_1| + |X_2| + \dots + |X_n| - a_n \} / b_n \xrightarrow{d} N(0, 1)$$

[Indian Civil Services, 1982]

12. Using C.L.T. show that

$$\lim_{n \rightarrow \infty} \left[ e^{-n} \sum_{k=0}^n \frac{n^k}{k!} \right] = \frac{1}{2} = \lim_{n \rightarrow \infty} \int_0^n \frac{e^{-t} \cdot t^{n-1}}{(n-1)!} dt$$

(Indian Civil Services, 1984)

13. Let  $\{X_n, n = 1, 2, \dots\}$  be a sequence of independent Bernoulli variates such that:  $P(X_n = 1) = p_n = 1 - P(X_n = 0), n = 1, 2, \dots, (q_n = 1 - p_n)$ .

Show that if  $\sum p_n q_n = \infty, (n = 1, 2, \dots, \infty)$ , then the CLT holds for the sequence  $\{X_n\}$ . What happens if  $\sum p_n q_n < \infty$ . (Indian Civil Services, 1988)

14. Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed r.v.'s with  $E(X_i) = \mu$ ;  $\text{Var}(X_i) = \sigma^2$ ;  $(0 < \sigma^2 < \infty)$ ;  $i = 1, 2, \dots, n$  and  $E(X_i - \mu)^4 = \sigma^4 + 1$ .

(a) State weak law of large numbers.

(b) If  $\Gamma \left[ \frac{1}{n} (X_1^2 + X_2^2 + \dots + X_n^2) - c \right] \rightarrow 0$  as  $n \rightarrow \infty$ , find  $c$ .

**Hint.** By Khinchines theorem  $c = E X_i^2 = \sigma^2 + \mu^2$  (finite).

(c) State the Lindeberg-Levy Central Limit theorem.

(d) Find  $\lim_{n \rightarrow \infty} P \left[ \sigma^2 - \frac{1}{\sqrt{n}} \leq \frac{(X_1 - \mu)^2 + \dots + (X_n - \mu)^2}{n} \leq \sigma^2 + \frac{1}{\sqrt{n}} \right]$

[Delhi Univ. B.A. (Stat. Hons.), Spl. Course 1986]

$$\begin{aligned} \text{Hint. } p_n &= P \left[ -\frac{1}{\sqrt{n}} \leq \frac{\sum (X_i - \mu)^2}{n} - \sigma^2 \leq \frac{1}{\sqrt{n}} \right] \\ &= P \left[ -\sqrt{n} \leq \sum (X_i - \mu)^2 - n \sigma^2 \leq \sqrt{n} \right] \\ &= P \left[ -1 \leq \sum_{i=1}^n [(X_i - \mu)^2 - \sigma^2] / \sqrt{n} \leq 1 \right] \\ &= P \left[ -1 \leq S_n / \sqrt{n} \leq 1 \right] \end{aligned} \quad \dots(*)$$

$$\text{where } S_n = \sum_{i=1}^n [(X_i - \mu)^2 - \sigma^2] = \sum_{i=1}^n U_i$$

$$\text{where } U_i = (X_i - \mu)^2 - \sigma^2, i = 1, 2, \dots, n; \text{ are i.i.d. r.v.'s.}$$

$$\Rightarrow E(U_i) = E((X_i - \mu)^2 - \sigma^2) = \sigma^2 - \sigma^2 = 0$$

$$\text{Var } U_i = \text{Var}[(X_i - \mu)^2 - \sigma^2] = \text{Var}(X_i - \mu)^2$$

$$= E(X_i - \mu)^4 - [E(X_i - \mu)^2]^2 = \sigma^4 + 1 - \sigma^4 = 1$$

$\vdash : E(X_i - \mu)^4 = \sigma^4 + 1$  (Given)

$$\therefore E(S_n) = \sum_{i=1}^n E(U_i) = 0; \text{Var}(S_n) = \text{Var}\left(\sum_{i=1}^n U_i\right) = \sum_{i=1}^n \text{Var}(U_i) = n$$

(As  $U_i$ 's are i.i.d.)

Hence by C.L.T.  $\frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n}{\sqrt{n}} \xrightarrow{L} N(0, 1)$  as  $n \rightarrow \infty$  ...(\*\*)

From (\*) and (\*\*), we get

$$\begin{aligned} \lim_{n \rightarrow \infty} p_n &= \lim_{n \rightarrow \infty} [-1 \leq S_n/\sqrt{n} \leq 1] = P(-1 \leq Z \leq 1), \text{ where } Z \sim N(0, 1) \\ &= 2 \times P(0 \leq Z \leq 1) = 2 \times 0.3413 = 0.6826 \end{aligned}$$

15. Let  $X_1, X_2, \dots, X_n$  be i.i.d.  $N(0, 1)$  variates. Show that the limiting distribution of

$$\sqrt{n}(X_1 + X_2 + \dots + X_n)/(X_1^2 + X_2^2 + \dots + X_n^2) \sim N(0, 1) \text{ as } n \rightarrow \infty.$$

**Hint.** Use Cramer's Theorem.

16. Let  $X_1, X_2, \dots, X_{2n}$  be i.i.d.  $N(0, 1)$  variates. Find the limiting distribution of  $Z_n = U_n/V_n$  where

$$U_n = \left( \frac{X_1}{X_2} + \frac{X_3}{X_4} + \dots + \frac{X_{2n-1}}{X_{2n}} \right), \quad V_n = X_1^2 + X_2^2 + \dots + X_n^2.$$

**Hint.**  $X_i$  are i.i.d.  $N(0, 1)$ ;  $i = 1, 2, \dots, 2n$ ;  $E(X_i^2) = \text{Var } X_i = 1$   
 $\Rightarrow (X_{2i-1}/X_{2i})$  are i.i.d. standard Cauchy variates;  $i = 1, 2, \dots, n$

$$\Rightarrow \lim_{n \rightarrow \infty} \frac{U_n}{n} \xrightarrow{L} \text{Standard Cauchy Variate} = C(0, 1),$$

(Being the mean of i.i.d. standard Cauchy variates)

$$\lim_{n \rightarrow \infty} \frac{V_n}{n} = \frac{X_1^2 + X_2^2 + \dots + X_n^2}{n} \xrightarrow{P} E X_i^2 = 1,$$

(By Khinchine's WLLN)

$$\therefore \frac{U_n}{V_n} = \left( \frac{U_n}{n} \right) / \left( \frac{V_n}{n} \right) \xrightarrow{L} C(0, 1)$$

(By Cramer's Theorem)

17. Let  $\{X_n\}$  be a sequence of i.i.d. r.v.'s with mean  $\alpha$  and variance  $\sigma^2$  and let  $\{Y_n\}$  be another sequence of i.i.d. r.v.'s with mean  $\beta$  ( $\neq 0$ ) and variance  $\sigma_1^2$ . Find the limiting distribution of :

$$Z_n = \sqrt{n}(\bar{X}_n - \alpha)/\bar{Y}_n \text{ where } \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$$

**Hint.**  $U_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sigma/\sqrt{n}} \xrightarrow{L} Z \sim N(0, 1)$  (By CLT)

$$V_n = \bar{Y}_n \xrightarrow{P} E(Y_n) = \beta$$

By Cramer's Theorem:

$$\lim_{n \rightarrow \infty} \frac{U_n}{V_n} = \frac{\sqrt{n}(\bar{X}_n - \alpha)}{\sigma \bar{Y}_n} \xrightarrow{L} \frac{Z}{\beta}, \text{ where } Z \sim N(0, 1)$$

$$\Rightarrow \lim_{n \rightarrow \infty} \frac{\sqrt{n}(\bar{X}_n - \alpha)}{\bar{Y}_n} \xrightarrow{L} \frac{\sigma}{\beta} Z \sim N\left(0, \frac{\sigma^2}{\beta^2}\right)$$

18. Let  $n$  numbers  $X_1, X_2, \dots, X_n$  in decimal form, be each approximated by the closest integer. If  $X_i$  is the  $i$ th true number and  $Y_i$  is the nearest integer, then  $U_i = X_i - Y_i$ , is the error made by the rounding process. Suppose that  $U_1, U_2, \dots, U_n$  are independent and each is uniform on  $(-0.5, 0.5)$ .

(i) What is the probability that the true sum is within ' $a$ ' units of the approximated sum?

(ii) If  $n = 300$  terms are added, find ' $a$ ' so that we are 95% sure that the approximation is within ' $a$ ' units of the true sum.

$$\text{HInt. Reqd. Prob. } p' = P\left[\left|\sum_{i=1}^n (X_i - Y_i)\right| \leq a\right] = P\left[-a \leq \sum_{i=1}^n U_i \leq a\right]$$

Now use Lindeberg Levy C.L.T. for i.i.d. r.v.'s

$U_i \sim U[-0.5, 0.5]$  with  $E(U_i) = 0$ ,  $\text{Var}(U_i) = 1/12$

Ans. (i)  $p = 2 \Phi(a \sqrt{12/n}) - 1$ ;

(ii)  $p = 0.95 \Rightarrow \Phi(a \sqrt{12/300}) = 0.975 \Rightarrow \frac{a}{\sqrt{12/300}} = 1.96 \Rightarrow a = 9.8$ .

8.11. Compound Distributions. Consider a random variable  $X$  whose distribution depends on a single parameter  $\theta$  which instead of being regarded as a fixed constant, is also a random variable following a particular distribution. In this case, we say that the random variable  $X$  has a *compound or composed distribution*.

8.11.1. Compound Binomial Distribution. Let us suppose that  $X_1, X_2, X_3, \dots$  are identically and independently distributed Bernoulli variates with

$$P(X_i = 1) = p \text{ and } P(X_i = 0) = q = 1 - p$$

For a fixed  $n$ , the random variable  $X = X_1 + X_2 + \dots + X_n$  is a Binomial variate with parameters  $n$  and  $p$  and probability function:

$$P(X = r) = \binom{n}{r} p^r q^{n-r}; r = 0, 1, 2, \dots, n$$

which gives the probability of  $r$  successes in  $n$  independent trials with constant probability ' $p$ ' of success for each trial.

Now suppose that  $n$ , instead of being regarded as a fixed constant, is also a random variable following Poisson law with parameter  $\lambda$ . Then

$$P(n = k) = \frac{e^{-\lambda} \lambda^k}{k!}; k = 0, 1, 2, \dots$$

In such a case  $X$  is said to have compound binomial distribution. The joint probability function of  $X$  and  $n$  is given by

$$P(X = r \cap n = k) = P(n = k) P(X = r | n = k)$$

$$= \frac{e^{-\lambda} \lambda^k}{k!} \binom{k}{r} p^r q^{k-r},$$

since  $P(X = r | n = k)$  is the probability of  $r$  successes in  $k$  trials. Obviously,  $r \leq k \Rightarrow k \geq r$ .

The marginal distribution of  $X$  is given by

$$\begin{aligned} P(X = r) &= \sum_{k=r}^{\infty} P(X = r \cap n = k) \\ &= e^{-\lambda} p^r \sum_{k=r}^{\infty} \binom{k}{r} \frac{\lambda^k q^{k-r}}{k!} = \frac{e^{-\lambda} (\lambda p)^r}{r!} \sum_{k=r}^{\infty} \frac{(\lambda q)^{k-r}}{(k-r)!} \\ &= \frac{e^{-\lambda} (\lambda p)^r}{r!} \sum_{j=0}^{\infty} \frac{(\lambda q)^j}{j!}, \quad (j = k-r) \\ &= \frac{e^{-\lambda} (\lambda p)^r}{r!} e^{\lambda q} = \frac{e^{-\lambda p} (\lambda p)^r}{r!}, \end{aligned}$$

which is the probability function of a Poisson variate with parameter  $\lambda p$ .

Hence  $E(X) = \lambda p$  and  $\text{Var}(X) = \lambda p$

We give below some of the practical situations where we would come across compound Binomial distribution:

1. Suppose that the probability of an insect laying  $n$  eggs is given by the Poisson distribution  $e^{-\lambda} \lambda^n / n!$  and the probability of an egg developing is  $p$ . Assuming natural independence of eggs, the probability of a total of  $k$  survivors is given by the Poisson distribution with parameter  $\lambda p$ .

2. The probability that a radioactive substance gives off  $n$  Beta particles in a unit of time is  $P(\lambda), (n = 0, 1, 2, \dots)$ . The probability that a given particle will strike a counter and be registered is  $p$ . Then the probability of registering  $n$  Beta particles in a unit of time is also  $P(\lambda p)$ .

3. If the probability of number of hits by lightning during any time interval  $t$  is  $P(\lambda t)$  and if the probability of its hitting and damaging an individual is  $p$  then (assuming stochastic independence) the total damage during time ' $t$ ' is  $P(\lambda t p)$ .

**8.11.2. Compound Poisson Distribution.** Let  $X$  be a  $P(\lambda)$  so that

$$P(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}; \quad r = 0, 1, 2, \dots$$

where  $\lambda$  itself is a continuous random variable with generalised gamma density

$$g(\lambda) = \begin{cases} \frac{a^v}{\Gamma(v)} e^{-a\lambda} \lambda^{v-1}; & \lambda \geq 0, a > 0, v > 0 \\ 0, & \lambda \leq 0 \end{cases}$$

Let us consider the two dimensional random vector  $(X, \lambda)$  in which one variable is discrete and the other is continuous. For a constant  $h > 0$  and  $\lambda_1 > 0$ , the joint density of  $X$  and  $\lambda$  is given by

$$P(X = r \cap \lambda_1 \leq \lambda \leq \lambda_1 + h) = P(\lambda_1 \leq \lambda \leq \lambda_1 + h) P(X = r | \lambda_1 \leq \lambda \leq \lambda_1 + h)$$

Dividing both sides by  $h$  and proceeding to the limits as  $h \rightarrow 0$ , we get

$$\lim_{h \rightarrow 0} \frac{P(X = r \cap \lambda_1 \leq \lambda \leq \lambda_1 + h)}{h} = \lim_{h \rightarrow 0} P(X = r | \lambda_1 \leq \lambda \leq \lambda_1 + h) \times \lim_{h \rightarrow 0} \frac{P(\lambda_1 \leq \lambda \leq \lambda_1 + h)}{h}$$

But

$$\lim_{h \rightarrow 0} \frac{P(\lambda_1 \leq \lambda \leq \lambda_1 + h)}{h} = \lim_{h \rightarrow 0} \frac{G(\lambda_1 + h) - G(\lambda_1)}{h} = G'(\lambda_1) = g(\lambda_1)$$

where  $G(\cdot)$  is the distribution function and  $g(\cdot)$  is the p.d.f. of  $\lambda$ .

$$\therefore \lim_{h \rightarrow 0} \frac{P(X = r \cap \lambda_1 \leq \lambda \leq \lambda_1 + h)}{h} = \frac{e^{-\lambda_1} \lambda_1^r}{r!} \cdot \frac{a^v}{\Gamma(v)} \lambda_1^{v-1} e^{-a \lambda_1}$$

Integrating w.r.to  $\lambda_1$  over 0 to  $\infty$  and using gamma integral, the marginal probability function of  $X$  is given by

$$\begin{aligned} P(X = r) &= \frac{a^v}{\Gamma(v) r!} \int_0^\infty e^{-(1+a)\lambda} \lambda^{r+v-1} d\lambda \\ &= \frac{a^v}{\Gamma(v) r!} \cdot \frac{\Gamma(r+v)}{(1+a)^{r+v}} \\ &= \left( \frac{a}{1+a} \right)^v \frac{v(v+1)(v+2)\dots(v+r-1)}{(1+a)^{r+v} r!} \\ &= \left( \frac{a}{1+a} \right)^v (-1)^r \binom{-v}{r} \left( \frac{1}{1+a} \right)^r \\ &= \binom{-v}{r} p^v (-q)^r ; r = 0, 1, 2, \dots \end{aligned}$$

where  $p = a/(1+a)$ ,  $q = 1-p = 1/(1+a)$

Thus the marginal distribution of  $X$  is a negative binomial with parameters  $(v, p)$ .

### EXERCISE 8(h)

1. (a) What do you mean by a compound distribution? Obtain the probability function of compound Poisson distribution and identify it.

(b) What is the Compound Binomial distribution? Obtain its probability function and identify the distribution.

2. If  $X$  is a random variable with p.d.f.,

$$f(x) = \frac{1}{\Gamma(n+1)} e^{-x} x^n, x \geq 0$$

where  $n$  is a positive integer, and the discrete random variable  $Y$  has a Poisson distribution with parameter  $\lambda$ , show that  $P(X \geq \lambda) = P(Y \leq n)$ .

**Hint.**  $P(X' \geq \lambda) = 1 - \frac{1}{\Gamma(n+1)} \int_0^\lambda e^{-x} x^n dx$

Integrating by parts successively, we get the result.

3. If  $X'$  has a Poisson distribution:

$$P(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}; r = 0, 1, 2, \dots$$

where the parameter  $\lambda$  is a random variable of the continuous type with the density function:  $f(\lambda) = \frac{a^v}{\Gamma(v)} \cdot e^{-a\lambda} \lambda^{(v-1)}; \lambda \geq 0, a > 0, v > 0,$

derive the distribution of  $X$ .

Show that the characteristic function of  $X$  is given by

$$\Phi_X(t) = E(e^{itX}) = q^v (1 - pe^{it})^{-v}, \text{ where } p = 1/(1+a), q = 1-p$$

[South Gujarat Univ. M.Sc. 1991]

4. The conditional distribution of a continuous random variable  $X$  for a discrete random variable  $Y$ , assuming a value  $n$  is

$$dF = n (1-x)^{n-1} dx, 0 \leq x \leq 1$$

The distribution of  $Y$  is:  $P(Y = n) = (\frac{1}{2})^n; n = 1, 2, 3, \dots$

Find the marginal distribution of  $X$ .

5. Given  $f(x|y) = \frac{e^{-y} y^x}{x!}$  and  $h(y) = e^{-y}$ , where  $X$  is discrete, i.e.,  $x = 0, 1, 2, \dots$  and  $Y$  is continuous,  $y \geq 0$ ; show that the marginal distribution of  $X$  is geometric, i.e.,  $g(x) = (\frac{1}{2})^{x+1}$ .

6. The conditional probability that the random variable  $X$  should lie within the range  $dx$  for a given  $\sigma$  is given by

$$\frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (\bar{x} - \mu)^2 / \sigma^2 \right\} dx, -\infty < x < \infty$$

while the probability of  $\sigma$  itself lying within the range  $d\sigma$  is

$$\frac{1}{\sigma_0^2} \exp \left\{ -\frac{1}{2} \sigma^2 / \sigma_0^2 \right\} \sigma d\sigma, 0 < \sigma < \infty$$

where  $\sigma_0$  is a constant. Show that the unconditional (i.e., marginal) distribution of  $X$  has the following probability function :

$$\frac{1}{2\sigma_0} \exp \left\{ -(1/\sigma_0) |x - \mu| \right\}, -\infty < x < \infty$$

7. Let  $X \sim U[0, 1]$  and  $Y | (X = x) \sim B(n; x)$  i.e.,

$$P(Y = y | X = x) = \binom{n}{y} x^y (1-x)^{n-y}, y = 0, 1, 2, \dots, n$$

Find the distribution of  $Y$ . Also find  $E(Y)$ .

Ans.  $P(Y = y) = 1/(n+1), y = 0, 1, \dots, n \Rightarrow Y \sim U[0, 1, 2, \dots, n]; E(Y) = n/2$

**8.12. Pearson's Distributions.** Given a set of observations from a population, the first question that arises in our mind is about the nature of the parent population. A vague idea is provided by the frequency polygon (or frequency curve) but the information is totally inadequate and unreliable, because the sample observations may not cover the entire range of the parent distribution. Moreover, an unusually high frequency in one class, arising out of sheer chance, may completely distort the shape of the frequency curve.

Consequently, to determine the frequency curve, we resort to the technique of curve-fitting to the given data. The failure of the normal distribution to fit many distributions which are observed in practice for continuous variables necessitated the development of generalised system of frequency curves. Since a trial and error approach is clearly undesirable, an elastic system of frequency curves must be evolved, which should incorporate, if not all, at least the most common of the distributions. *Pearsonian system of frequency curves* is one of the most important approaches in this direction, in which we decide about the shape of the curve on the basis of a 'criterion  $\kappa$ ' calculated from the sample observations.

Karl Pearson's first memoir dealing with generalised 'frequency' curves appeared in 1895. In this paper and the subsequent two papers published in 1908 and 1916, Karl Pearson developed a set of frequency curves which could be obtained by assigning values to the parameters in a certain first order differential equation.

**Genesis of Pearson's Frequency Curves.** Experience tells us that most of the frequency distributions possess the following obvious and common characteristic:

"They rise from a low frequency to a maximum frequency and then again fall to the low frequency as the variable  $X$  increases. This suggests a unimodal frequency curve  $y = f(x)$  with high contact at the extremities of the range, i.e.,  $\frac{dy}{dx} = 0$  when  $y = 0$ . Accordingly, Karl Pearson proposed the following differential equation for the frequency curve  $y = f(x)$ ,

$$\frac{dy}{dx} = \frac{y(x-a)}{F(x)} \quad \dots(8.32)$$

where  $F(x)$  is an arbitrary function of  $x$  not vanishing at  $x = a$ , the mode of the distribution. Expanding  $F(x)$  by Maclaurin's theorem, we get  $F(x) = b_0 + b_1 x + b_2 x^2 + \dots$  and retaining only the first three terms we get the differential equation of the Pearsonian system of frequency curves as

$$\frac{dy}{dx} = \frac{y(x-a)}{b_0 + b_1 x + b_2 x^2} \Rightarrow \frac{df(x)}{dx} = f'(x) = \frac{(x-a)f(x)}{b_0 + b_1 x + b_2 x^2} \dots(8.32a)$$

where  $a, b_0, b_1$  and  $b_2$  are the constants to be calculated from the given data.

**Remark.** Equation (8.32a) can also be obtained as a limiting case of Hyper-geometric distribution (c.f. Advanced statistics Vol. II by Kendall).

**8.12.1. Determination of the Constants of the Equation in Terms of Moments.** Multiplying both sides of (8.32 a) by  $x^n$  for integral  $n > 0$  and integrating over the entire range of the variable  $X$  say  $(\alpha, \beta)$ , we get

$$\int_{\alpha}^{\beta} x^n (b_0 + b_1 x + b_2 x^2) f'(x) dx = \int_{\alpha}^{\beta} x^n (x - a) f(x) dx$$

$$\Rightarrow \left[ x^n (b_0 + b_1 x + b_2 x^2) f(x) \right]_{\alpha}^{\beta} - \int_{\alpha}^{\beta} \left[ nb_0 x^{n-1} + (n+1)b_1 x^n + (n+2)b_2 x^{n+1} \right] f(x) dx$$

$$= \int_{\alpha}^{\beta} x^{n+1} f(x) dx - a \int_{\alpha}^{\beta} x^n f(x) dx$$

Assuming high order contact at the extremities so that

$$\left| x^n f(x) \right|_{\alpha}^{\beta} = 0 \text{ i.e., } x^n f(x) \rightarrow 0 \text{ as } x \rightarrow \alpha \text{ or } \beta, \text{ we get} \quad \dots(*)$$

$$- [nb_0 \mu_{n-1} + (n+1)b_1 \mu_n + (n+2)b_2 \mu_{n+1}] = \mu_{n+1} - a \mu_n$$

(assuming that  $X$  is measured from mean and this we can do without any loss of generality). Thus the recurrence relation between the moments becomes

$$nb_0 \mu_{n-1} + [(n+1)b_1 - a] \mu_n + [(n+2)b_2 + 1] \mu_{n+1} = 0, \quad \dots(**)$$

$$n = 1, 2, 3, \dots$$

Integrating (8.32 a) w.r.to  $x$  within the limits  $(\alpha, \beta)$  and using (\*), we get  

$$(b_1 - a) + (2b_2 + 1) = 0 \quad \dots(***)$$

Putting  $n = 1, 2$ , and  $3$  in  $(**)$  and solving these equations and  $(***)$  with the help of determinants and using  $\mu_0 = 1, \mu_1 = 0$ , we get

$$b_0 = -\frac{\mu_2(4\mu_2\mu_4 - 3\mu_3^2)}{2(5\mu_2\mu_4 - 9\mu_2^3 - 6\mu_3^2)} = -\frac{\sigma^2(4\beta_2 - 3\beta_1)}{2(5\beta_2 - 6\beta_1 - 9)} \quad \dots(8.33)$$

$$a = b_1 = -\frac{\mu_3(\mu_4 + 3\mu_2^2)}{2(5\mu_2\mu_4 - 9\mu_2^3 - 6\mu_3^2)} = -\frac{\sigma\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

$$b_2 = -\frac{(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3)}{2(5\mu_2\mu_4 - 9\mu_2^3 - 6\mu_3^2)} = -\frac{(2\beta_2 - 3\beta_1 - 6)}{2(5\beta_2 - 6\beta_1 - 9)}$$

where  $\mu_2 = \sigma^2$ ,  $\beta_1 = \mu_3^2/\mu_2^2$  and  $\beta_2 = \mu_4/\mu_2^2$ .

Thus the Pearson's system (8.32 a) is completely specified by the first four moments.

### 8.12.2. Pearson Measure of Skewness.

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{0 - a}{\sqrt{\mu_2}}$$

$$= \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} \quad \dots(8.34)$$

**8.12.3. "Criterion  $\kappa$ ".** Equation (8.32 a) can be re-written as:

$$\frac{df}{f} = \frac{(x-a) dx}{b_0 + b_1 x + b_2 x^2}, f=f(x)$$

Integrating we get

$$\log(f/c) = \int \frac{(x-a)}{b_0 + b_1 x + b_2 x^2} dx = I \text{ (say)},$$

where  $c$  is the constant of integration.

$$\therefore f(x) = c \exp [I]$$

Thus  $f$  depends on  $I$ , which further depends on the roots of the equation

$$b_0 + b_1 x + b_2 x^2 = 0 \quad \dots(8.35)$$

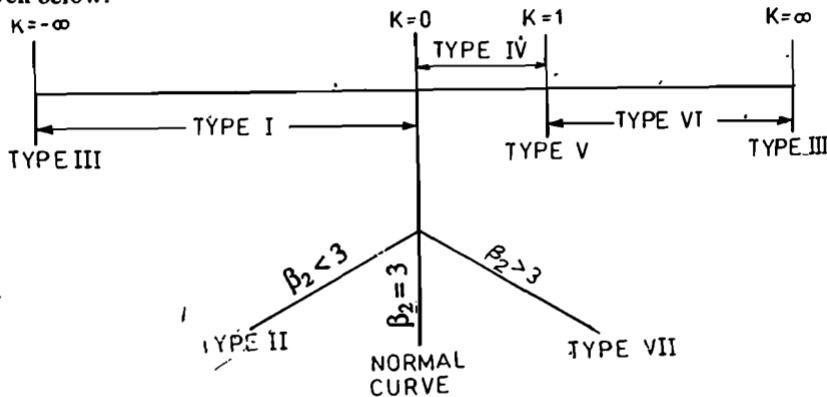
Now,

$$\begin{aligned} b_0 + b_1 x + b_2 x^2 &= b_2 \left[ x^2 + \frac{b_1}{b_2} x + \frac{b_0}{b_2} \right] \\ &= b_2 \left[ x - \frac{-b_1 + \sqrt{b_1^2 - 4 b_0 b_2}}{2 b_2} \right] \left[ x - \frac{-b_1 - \sqrt{b_1^2 - 4 b_0 b_2}}{2 b_2} \right] \\ &= b_2 \left[ x + \frac{b_1}{2 b_2} - \frac{\sqrt{b_1^2 - 4 b_0 b_2}}{\sqrt{4 b_2^2}} \right] \left[ x + \frac{b_1}{2 b_2} + \frac{\sqrt{b_1^2 - 4 b_0 b_2}}{\sqrt{4 b_2^2}} \right] \\ &= b_2 \left[ x + \frac{b_1}{2 b_2} - \sqrt{\frac{b_0}{b_2} (\kappa - 1)} \right] \left[ x + \frac{b_1}{2 b_2} + \sqrt{\frac{b_0}{b_2} (\kappa - 1)} \right] \end{aligned}$$

$$\text{where } \kappa = b_1^2 / (4 b_0 b_2), \quad \dots(8.35)$$

determines the criterion for obtaining the form of the frequency curve.

A brief description of various Pearsonian curves for different values of  $\kappa$  is given below:



**8.12.4. Pearson's Main Type 1.** This curve is obtained when the roots of the quadratic equation are real and of opposite sign, i.e., when  $\kappa < 0$ .

Shifting the origin to the mode  $x = a$ , the equation becomes

$$\frac{df}{dx} = \frac{xf}{B_0 + B_1 x + B_2 x^2} = \frac{xf}{B_2(x+\alpha)(x-\beta)}$$

where  $B_0 = b_0$ ,  $B_1 = b_1$  and  $B_2 = b_2$ .

$$\Rightarrow \frac{d}{dx} (\log f) = \frac{1}{B_2(\alpha+\beta)} \left[ \frac{\alpha}{x+\alpha} + \frac{\beta}{x-\beta} \right]$$

Integrating both sides w.r.to.  $x$ , we get

$$\log f = \log(x+\alpha)^{\alpha/B_2(\alpha+\beta)} + \log(x-\beta)^{\beta/B_2(\alpha+\beta)} + \log C$$

$$\therefore f = C(x+\alpha)^{\alpha/B_2(\alpha+\beta)}(x-\beta)^{\beta/B_2(\alpha+\beta)}$$

$$\Rightarrow f = y_0 \left( 1 + \frac{x}{\alpha} \right)^{\alpha/B_2(\alpha+\beta)} \left( 1 - \frac{x}{\beta} \right)^{\beta/B_2(\alpha+\beta)}, -\alpha \leq x \leq \beta \quad \dots(*)$$

Let  $\alpha = a_1$ ,  $\beta = a_2$ ,  $m_1 = \frac{\alpha}{B_2(\alpha+\beta)}$  and  $m_2 = \frac{\beta}{B_2(\alpha+\beta)}$  so that

$$\frac{m_1}{a_1} = \frac{m_2}{a_2} = \frac{1}{B_2(a_1+a_2)}, \text{ then } (*) \text{ may be written as}$$

$$f(x) = y_0 \left( 1 + \frac{x}{a_1} \right)^{m_1} \left( 1 - \frac{x}{a_2} \right)^{m_2}, -a_1 \leq x \leq a_2 \quad \dots(8-36)$$

which is a standard form of Type 1.

*Determination of  $y_0$ :*

$$1 = y_0 \int_{-a_1}^{a_2} \left( 1 + \frac{x}{a_1} \right)^{m_1} \left( 1 - \frac{x}{a_2} \right)^{m_2} dx$$

Put  $x = (a_1 + a_2)z - a_1$  so that  $dx = (a_1 + a_2)dz$

$$\Rightarrow 1 = y_0 \int_0^1 \frac{(a_1 + a_2)^{m_1}}{a_1^{m_1}} z^{m_1} \frac{(a_1 + a_2)^{m_2}}{a_2^{m_2}} (1-z)^{m_2} (a_1 + a_2) dz$$

$$\Rightarrow 1 = y_0 \left[ \frac{(a_1 + a_2)^{m_1 + m_2 + 1}}{a_1^{m_1} a_2^{m_2}} \right] B(m_1 + 1, m_2 + 1)$$

$$\therefore y_0 = \frac{a_1^{m_1} a_2^{m_2}}{(a_1 + a_2)^{m_1 + m_2 + 1} B(m_1 + 1, m_2 + 1)}$$

**Remark.** It may be noted that Beta distribution is a particular case of Type I distribution.

*Determination of moments:*

$$\begin{aligned} \mu_n' &= y_0 \int_{-a_1}^{a_2} (x + a_1)^n \left( 1 + \frac{x}{a_1} \right)^{m_1} \left( 1 - \frac{x}{a_2} \right)^{m_2} dx \\ &= y_0 \frac{(a_1 + a_2)^{m_1 + m_2 + n + 1}}{a_1^{m_1} a_2^{m_2}} B(n + m_1 + 1, m_2 + 1), \end{aligned}$$

where  $x = (a_1 + a_2)z - a_1$

$$= \frac{(a_1 + a_2)^n}{B(m_1 + 1, m_2 + 1)} \cdot B(n + m_1 + 1, m_2 + 1)$$

[On simplification]

**8.12.5. Pearson's Type IV.** This curve is obtained when the roots are imaginary or when

$$B_1^2 < 4 B_0 B_2, \text{ i.e., } 0 < \kappa < 1$$

$$\frac{1}{f} \cdot \frac{df}{dx} = \frac{x}{B_0 + B_1 x + B_2 x^2}$$

[Origin at mode]

$$\Rightarrow \frac{d}{dx} (\log f) = \frac{x}{B_2 \left[ \left( x + \frac{B_1}{2B_2} \right)^2 + \left( \frac{B_0}{B_2} - \frac{B_1^2}{4B_2^2} \right) \right]} \\ = \frac{(x + \gamma) - \gamma}{B_2 [(x + \gamma)^2 + \delta^2]} = \frac{2(x + \gamma)}{2B_2 [(x + \gamma)^2 + \delta^2]} - \frac{2\gamma}{2B_2 [(x + \gamma)^2 + \delta^2]}$$

$$\log f = \frac{1}{2B_2} \log [(x + \gamma)^2 + \delta^2] - \frac{\gamma}{B_2} \cdot \frac{1}{\delta} \tan^{-1} \frac{x + \gamma}{\delta} + \log C$$

$$\therefore f = C \left[ (x + \gamma)^2 + \delta^2 \right]^{\frac{1}{2B_2}} \exp \left\{ -\frac{\gamma}{B_2 \delta} \tan^{-1} \frac{x + \gamma}{\delta} \right\}$$

$$\text{Put } \frac{x + \gamma}{\delta} = \frac{x}{a} \text{ and } \frac{\gamma}{B_2 \delta} = v$$

$$\text{Hence } f(x) = y_0 \left( 1 + \frac{x^2}{a^2} \right)^{-m} e^{-v \tan^{-1}(x/a)} ; -\infty < x < \infty, (m, v) > 0 \quad \dots(8.37)$$

which is a standard form of Type IV with origin at  $\left( -\frac{B_1}{2B_2}, 0 \right)$ . The curve is skew and has unlimited range in both the directions.

*Determination of  $y_0$ :*

$$1 = y_0 \int_{-\infty}^{\infty} \left( 1 + \frac{x^2}{a^2} \right)^{-m} e^{-v \tan^{-1}(x/a)} dx \quad [\text{Put } x = a \tan \theta]$$

$$= ay_0 \int_{-\pi/2}^{\pi/2} \cos^{2m-2} \theta e^{-v \theta} d\theta = ay_0 F(2m-2, v)$$

$$\therefore y_0 = \frac{1}{a F(2m-2, v)}$$

$$\text{Hence } f(x) = \frac{1}{a F(2m-2, v)} \left( 1 + \frac{x^2}{a^2} \right)^{-m} e^{-v \tan^{-1}(x/a)}$$

**8-12-6. Pearson's Type VI.** The curve is obtained when the roots are real and are of the same sign. This is obtained when  $B_0, B_2$  are of the same sign or, in other words, when  $\kappa > 1$ .

Let  $-\alpha_1$  and  $-\alpha_2$  be the roots of the quadratic equation. Then

$$\begin{aligned}\frac{d}{dx}(\log f) &= \frac{x}{B_0 + B_1 x + B_2 x^2} = \frac{x}{B_2(x + \alpha_1)(x + \alpha_2)} \\ &= \frac{\alpha_1}{B_2(\alpha_1 - \alpha_2)} \cdot \frac{1}{(x + \alpha_1)} - \frac{\alpha_2}{B_2(\alpha_1 - \alpha_2)} \cdot \frac{1}{(x + \alpha_2)}\end{aligned}$$

$$\therefore \log f = \frac{\alpha_1}{B_2(\alpha_1 - \alpha_2)} \log(x + \alpha_1) - \frac{\alpha_2}{B_2(\alpha_1 - \alpha_2)} \log(x + \alpha_2) + \log C$$

$$\Rightarrow f = C(x + \alpha_1)^{\alpha_1/B_2(\alpha_1 - \alpha_2)} \cdot (x + \alpha_2)^{-\alpha_2/B_2(\alpha_1 - \alpha_2)}$$

Hence the probability density function is:

$$f(x) = y_0 \left(1 + \frac{x}{a_1}\right)^{m_1} \left(1 + \frac{x}{a_2}\right)^{-m_2} \quad \dots(8.38)$$

where  $a_1 = \alpha_1$ ,  $a_2 = \alpha_2$  and  $\frac{m_1}{a_1} = -\frac{m_2}{a_2}$ ;  $a_1, a_2 > 0$

This equation can also be written (on shifting the origin to  $-\alpha_2$  or  $-a_2$ ) as

$$f(x) = y_0 (x - a)^{q_2} x^{-q_1}, \quad a \leq x < \infty \quad \dots(8.38a)$$

$$\text{where } q_1 = \frac{\alpha_2}{B_2(\alpha_1 - \alpha_2)}, \quad q_2 = \frac{\alpha_1}{B_2(\alpha_1 - \alpha_2)}, \quad a = -\alpha_1$$

**Remark.** The curve is bell shaped if  $q_2 > 0$  and J-shaped if  $q_2 < 0$ .

*Determination of  $y_0$ :*

$$\begin{aligned}1 &= y_0 \int_a^\infty (x - a)^{q_2} x^{-q_1} dx \quad \left[ \text{Put } x = \frac{a}{1-z} \right] \\ &= y_0 \int_0^1 \left(\frac{a}{1-z}\right)^{-q_1} \left[\frac{a}{1-z} - a\right]^{q_2} \frac{a}{(1-z)^2} dz \\ &= y_0 a^{q_2 - q_1 + 1} \int_0^1 z^{q_2} \frac{1}{(1-z)^{q_2 - q_1 + 2}} dz \\ \therefore y_0 &= \frac{a^{q_1 - q_2 - 1}}{B(q_2 + 1, q_1 - q_2 - 1)}\end{aligned}$$

$$\text{Hence } f(x) = \frac{a^{q_1 - q_2 - 1}}{B(q_2 + 1, q_1 - q_2 - 1)} x^{-q_1} (x - a)^{q_2}, \quad a \leq x < \infty$$

The above discussion covers almost the whole range of  $\kappa$  but in limiting cases we get simple cases. The following are more important of the transition curves when one of the main type changes into another.

**8-12-7. Type III.** This is a transition type curve and is obtained when  $B_2 \neq 0, B_1 \neq 0$  or  $\kappa \rightarrow \pm \infty$ .

$$\begin{aligned}
 \frac{d}{dx} (\log f) &= \frac{x}{B_0 + B_1 x}, \text{ (origin is at mode).} \\
 &= \frac{B_1 x + B_0 - B_0}{B_1 (B_0 + B_1 x)} = \frac{1}{B_1} - \frac{B_0}{B_1 (B_0 + B_1 x)} \\
 \therefore \log f &= \frac{x}{B_1} - \frac{B_0}{B_1^2} \log (B_0 + B_1 x) + \text{const.} \\
 &= \text{const. } e^{x/B_1} (B_0 + B_1 x)^{-B_0/B_1^2} \\
 f(x) &= y_0 \left( 1 + \frac{x}{a} \right)^p e^{-px/a}; -a \leq x < \infty, \\
 \text{where } \frac{B_0}{B_1} &= a \text{ and } \frac{B_0}{B_1^2} = p
 \end{aligned} \tag{8.39}$$

This gives the Type III curve with origin at mode. The curve is usually bell shaped but becomes J-shaped when  $\beta_1 > 4$ .

**Remark.** The distribution can be transformed into the gamma form by using the transformation  $y = \frac{p}{a}(x + a)$ , when the curve reduces to

$$f(y) = \frac{1}{\Gamma(p+1)} e^{-y} y^p, 0 \leq y < \infty$$

**8.12.8. Type V.** This transition type is obtained when the roots are equal, i.e., when  $B_1^2 = 4B_0B_2$  or  $\kappa = 1$ .

$$\begin{aligned}
 \frac{d}{dx} (\log f) &= \frac{x}{B_2 \left[ \left( x + \frac{B_1}{2B_2} \right)^2 \right]} = \frac{2 \left[ x + \frac{B_1}{2B_2} \right] - \frac{B_1}{B_2}}{2B_2 \left[ \left( x + \frac{B_1}{2B_2} \right)^2 \right]} \\
 \therefore \log f &= \frac{1}{2B_2} \log \left( x + \frac{B_1}{2B_2} \right) + \frac{B_1}{2B_2^2} \frac{1}{\left[ x + \frac{B_1}{2B_2} \right]} + \text{const.} \\
 \Rightarrow f &= \text{const.} \left( x + \frac{B_1}{2B_2} \right)^{\frac{1}{B_2}} \exp \left[ \frac{B_1}{2B_2^2} \left( x + \frac{B_1}{2B_2} \right)^{-1} \right] \\
 \therefore f(x) &= y_0 X^{-p} e^{-q/X}, 0 \leq X < \infty
 \end{aligned} \tag{8.40}$$

$$\text{where } X = \left( x + \frac{B_1}{2B_2} \right), \frac{B_1}{2B_2^2} = -q \text{ and } \frac{1}{B_2} = -p.$$

**8.12.9. Type II.** This curve is obtained when  $B_1 = 0$  and  $B_0, B_2$  are of opposite sign, i.e.,  $\kappa = 0$ . The equation to the curve is

$$f(x) = y_0 \left[ 1 + \frac{x^2}{a^2} \right]^m, -a \leq x \leq a; \tag{8.41}$$

$$\text{where } m = \frac{1}{2B_2} > 0, a^2 = \frac{B_0}{B_2}$$

with origin at mean (mode).

**8-12-10. Type VII.** This curve is obtained when  $B_1 = 0$  and  $B_0, B_2$  are of the same sign, i.e.,  $\kappa = 0$  and  $B_0 B_2 > 0$ . The equation to the curve is

$$f(x) = y_0 \left[ 1 + \frac{x^2}{a^2} \right]^{-m}, -\infty < x < \infty \quad \dots(8-42)$$

$$\text{where } a^2 = \frac{B_0}{B_2} \quad \text{and} \quad m = -\frac{1}{2B_2}$$

with origin being at the mean (mode). This curve is usually bell shaped, symmetrical and of unlimited range in both the directions.

**8-12-11. Zero Type (Normal curve).** When  $B_1 = B_2 = 0$ , (18-33) implies that  $\beta_1 = 0$ , and  $\beta_2 = 3$  and we have

$$\frac{d}{dx} (\log f) = \frac{x}{B_0} \Rightarrow \log f = \frac{x^2}{2B_0} + \log C,$$

where  $C$  is the constant of integration.

$$\therefore f = C \exp(x^2/2B_0) = C \exp(-x^2/2\sigma^2), -\infty < x < \infty \quad \dots(8-43)$$

where  $B_0 = -\sigma^2$  and the origin is at mean. This is the normal distribution with mean zero and variance  $\sigma^2$ .

**8-12-12. Type VIII.** When  $B_0 = 0, B_1 > 0$ ,

$$f(x) = \frac{1-m}{a} \left( 1 + \frac{x}{a} \right)^m, -a \leq x \leq 0 \quad \dots(8-44)$$

**Type IX.** When  $B_0 = 0, B_1 < 0$  and  $\kappa < 0$

$$f(x) = \frac{1+m}{a} \left( 1 + \frac{x}{a} \right)^m, -a \leq x \leq 0 \quad \dots(8-45)$$

**Type X.** When  $B_0 = 0$  and  $B_2 = 0$ ,

$$f(x) = \frac{1}{\sigma} e^{-x/\sigma}, 0 \leq x < \infty, \sigma > 0 \quad \dots(8-46)$$

This is the p.d.f. of simple exponential distribution with parameter  $\sigma > 0$ .

**Type XI.** When  $B_0 = B_1 = 0$ , and  $\kappa > 1$

$$f(x) = b^{m-1} (m-1) x^{m-1}, b \leq x < \infty$$

**Type XII.** When  $5\beta_2 - 6\beta_1 - 9 = 0, \kappa < 0$

$$f(x) = \left( \frac{a_1}{a_2} \right)^m \frac{1}{(a_1 + a_2) B(1+m, 1-m)} \cdot \frac{\left( 1 + \frac{x}{a_1} \right)^m}{\left( 1 - \frac{x}{a_2} \right)^m}, a_1 \leq x \leq a_2$$

**Example 8.47.** Show that for a Pearson distribution :

$$\frac{df}{f} = \frac{(a+x)dx}{b_0 + b_1x + b_2x^2},$$

the characteristic function  $\varphi$  obeys the relation :

$$b_2\theta \frac{d^2\varphi}{d\theta^2} + (1+2b_2+b_1\theta) \frac{d\varphi}{d\theta} + (a+b_1+b_0\theta)\varphi = 0, \text{ where } \theta = it.$$

Deduce the recurrence relation for moments.

Show also that the cumulant generating function  $\psi$  obeys the relation :

$$b_2\theta \left\{ \frac{d^2\psi}{d\theta^2} + \left( \frac{d\psi}{d\theta} \right)^2 \right\} + (1+2b_2+b_1\theta) \frac{d\psi}{d\theta} + (a+b_1+b_0\theta) = 0,$$

Hence show that the cumulants obey the recurrence relation :

$$\left\{ 1 + (r+2)b_2 \right\} \kappa_{r+1} + rb_1 \kappa_r + rb_2 \left\{ \binom{r-1}{1} \kappa_2 \kappa_{r-1} + \binom{r-1}{2} \kappa_3 \kappa_{r-2} \right. \\ \left. + \dots + \binom{r-1}{j} (\kappa_{j+1} \kappa_{r-j} + \dots + \binom{r-1}{r-2} \kappa_{r-1} \kappa_2) \right\} = 0$$

$$\text{Solution. } (b_0 + b_1x + b_2x^2) \frac{df}{dx} = (a+x)f$$

$$\Rightarrow e^{\theta x} (b_0 + b_1x + b_2x^2) \frac{df}{dx} = e^{\theta x} (a+x)f$$

Integrating w.r.to.  $x$ , for the total range of  $x$ , assuming that integrals vanish at either limit, we get

$$\int_{-\infty}^{\infty} e^{\theta x} (b_0 + b_1x + b_2x^2) \frac{df}{dx} dx = \int_{-\infty}^{\infty} e^{\theta x} (a+x) f dx \\ \Rightarrow \left[ e^{\theta x} (b_0 + b_1x + b_2x^2) f \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \left\{ \theta e^{\theta x} (b_0 + b_1x + b_2x^2) \right. \\ \left. + e^{\theta x} (b_1 + 2b_2x) \right\} f dx = \int_{-\infty}^{\infty} e^{\theta x} (a+x) f dx \\ \Rightarrow 0 - \theta \left[ b_0 \varphi + b_1 \frac{d\varphi}{d\theta} + b_2 \frac{d^2\varphi}{d\theta^2} \right] - \left[ b_1 \varphi + 2b_2 \frac{d\varphi}{d\theta} \right] = \left( a \varphi + \frac{d\varphi}{d\theta} \right) *$$

$$\therefore \varphi = E(e^{itx}) = \int_{-\infty}^{\infty} e^{itx} f dx = \int_{-\infty}^{\infty} e^{\theta x} f dx,$$

[  $\because it = \theta$  ]

$$\frac{d\varphi}{d\theta} = \int_{-\infty}^{\infty} x e^{\theta x} f dx \text{ and } \frac{d^2\varphi}{d\theta^2} = \int_{-\infty}^{\infty} x^2 e^{\theta x} f dx$$

assuming differentiation is valid under integral sign.

$$\Rightarrow b_2 \theta \frac{d^2 \varphi}{d \theta^2} + (1 + 2b_2 + b_1 \theta) \frac{d \varphi}{d \theta} + (a + b_1 + b_0 \theta) \varphi = 0 \text{ (On simplification)} \quad \dots(1)$$

Differentiating  $n$  times w.r.t.  $\theta$ , using Leibnitz Theorem, we get

$$b_2 \left[ \theta \frac{d^{n+2} \varphi}{d \theta^{n+2}} + n \cdot \frac{d^{n+1} \varphi}{d \theta^{n+1}} \cdot 1 \right] + \left[ \left\{ \frac{d^{n+1} \varphi}{d \theta^{n+1}} (1 + 2b_2 + b_1 \theta) + \frac{d^n \varphi}{d \theta^n} \cdot nb_1 \right\} \right. \\ \left. + \left[ \left\{ \frac{d^n \varphi}{d \theta^n} (a + b_1 + b_0 \theta) + \left\{ \frac{d^{n-1} \varphi}{d \theta^{n-1}} \cdot nb_0 \right\} \right\} \right] = 0 \quad \dots(2)$$

Putting  $\theta = 0$  and using the relation  $\left[ \frac{d^n \varphi}{d \theta^n} \right]_{\theta=0} = \mu_n'$ , we get

$$nb_2 \mu'_{n+1} + (2b_2 + 1) \mu'_{n+1} + nb_1 \mu_n' + (b_1 + a) \mu_n' + nb_0 \mu'_{n-1} = 0$$

Shifting the origin to the mean, we get

$$[(n+2)b_2 + 1] \mu_{n+1} + [(n+1)b_1 + a] \mu_n + nb_0 \mu_{n-1} = 0 \quad \dots(3)$$

$$\text{Now } \varphi = e^\psi, \frac{d \varphi}{d \theta} = e^\psi \frac{d \psi}{d \theta} \text{ and } \frac{d^2 \varphi}{d \theta^2} = e^\psi \left[ \frac{d^2 \psi}{d \theta^2} + \left( \frac{d \psi}{d \theta} \right)^2 \right] \\ (\because \psi = \log \varphi)$$

Substituting these values in (1) and on simplification, we get

$$b_2 \theta \left[ \frac{d^2 \psi}{d \theta^2} + \left( \frac{d \psi}{d \theta} \right)^2 \right] + (1 + 2b_2 + b_1 \theta) \frac{d \psi}{d \theta} + (a + b_1 + b_0 \theta) = 0 \quad \dots(4)$$

Differentiating (4)  $r$  times w.r.t.  $\theta$  using Leibnitz Theorem, we get

$$b_2 \theta \left[ \frac{d^{r+2} \psi}{d \theta^{r+2}} + \frac{d^r}{d \theta^r} \left( \frac{d \psi}{d \theta} \right)^2 \right] + \binom{r}{1} b_2 \left[ \frac{d^{r+1} \psi}{d \theta^{r+1}} + \frac{d^{r-1}}{d \theta^{r-1}} \left( \frac{d \psi}{d \theta} \right)^2 \right] \\ + (1 + 2b_2 + b_1 \theta) \frac{d^{r+1} \psi}{d \theta^{r+1}} + \binom{r}{1} b_1 \frac{d^r \psi}{d \theta^r} = 0$$

$$\Rightarrow b_2 \theta \left[ \frac{d^{r+2} \psi}{d \theta^{r+2}} + \left\{ \frac{d^{r-1}}{d \theta^{r-1}} \left[ 2 \frac{d \psi}{d \theta} \cdot \frac{d^2 \psi}{d \theta^2} \right] \right\} \right] \\ + rb_2 \left[ \frac{d^{r+1} \psi}{d \theta^{r+1}} + \frac{d^{r-2}}{d \theta^{r-2}} \left\{ 2 \frac{d \psi}{d \theta} \cdot \frac{d^2 \psi}{d \theta^2} \right\} \right] \\ + (1 + 2b_2 + b_1 \theta) \frac{d^{r+1} \psi}{d \theta^{r+1}} + rb_1 \frac{d^r \psi}{d \theta^r} = 0$$

Putting  $\theta = 0$  and using the relation

$$\left[ \frac{d^n \Psi}{d \theta^n} \right]_{\theta=0} = K_n, \text{ we get}$$

$$\begin{aligned} & \{1 + (r+2)b_2\} K_{r+1} + \binom{r}{1} b_1 K_r + r b_2 \left\{ \binom{r-1}{1} K_2 \cdot K_{r-1} \right. \\ & \left. \cdot + \binom{r-1}{2} K_3 K_{r-2} + \dots + \binom{r-1}{r-2} K_{r-1} K_2 \right\} = 0 \quad (\text{On simplification}) \end{aligned}$$

### EXERCISE 8(i)

#### 2. Derive the differential equation

$$\frac{1}{y} \cdot \frac{dy}{dx} = \frac{x+a}{b_0 + b_1 x + b_2 x^2}$$

as the limiting form of the hypergeometric distribution.

Show that, for the Pearsonian family of distributions :

$$\frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{\sqrt{\beta_1} (\beta_2 + 3)}{(5 \beta_2 - 6 \beta_1 - 9)}$$

2. (a) State the differential equation for the Pearsonian system of curves and obtain the expressions for the constants in terms of moments. Obtain Type 1 distribution as a particular case of Pearson's system of frequency curves and describe method of fitting it by moments.

(b) Describe the procedure for classifying the Pearson family of distributions into various types. Show that all Pearsonian distributions are determined by the first four moments.

Show that 'Normal', 'Beta' and 'Gamma' distributions belong to the Pearson family.

(c) Assign the following distribution to one of the Pearson's types. Give the reasons for your answers

$$(i) \quad dF = K e^{-x^2/2} (x^2)^{(n/2)-1} dx^2, \quad 0 < x^2 < \infty$$

$$(ii) \quad dF = K \left( 1 + \frac{t^2}{n} \right)^{\left( \frac{-n-1}{2} \right)} dt, \quad -\infty < t < \infty.$$

3. What are the reasons for the adoption of the following general form to describe the Pearsonian system of frequency curves

$$\frac{d}{dx} f(x) = \frac{(x-a)f(x)}{b_0 + b_1 x + b_2 x^2} ?$$

Show that the Pearsonian curves can be characterised by a single criterion  $K$ . Outline the various types of curves for different values of  $K$ .

4. Obtain Pearson Type III curve in its usual form with node as origin, from the basic differential equation of the Pearsonian system of curves and establish a method of fitting this curve to the given data by the method of moments.

Hence otherwise, show that for this distribution  $2 \beta_2 = 3 (\beta_1 + 2)$ .

5. Derive the Beta distribution as a special case of the Pearsonian system of frequency functions expressed by

$$\frac{d(\log f)}{dx} = \frac{x+a}{b_0 + b_1 x + b_2 x^2}$$

6. (a) Derive Type 1 Pearsonian frequency curve and examine if the distribution given by

$$dP = y_0 (1 - y^2)^{(n+4)/2} dy, -1 \leq y \leq 1$$

reduces to that distribution.

(b) Express the constants  $y_0$ ,  $a$  and  $m$  of the distribution:

$$f(x) = y_0 \left( 1 - \frac{x^2}{a^2} \right)^m, -a < x < a$$

in terms of its  $\mu_2$  and  $\beta_2$ .

(c) Show that normal, gamma and beta distributions belong to the Pearsonian system.

7. Show that the following are members of the Pearson's system of curves and sketch them for some typical values of the constants.

$$(i) f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2} \cdot \frac{x^2}{\sigma^2}\right), -\infty < x < \infty$$

$$(iii) f(x) = \frac{1}{a B(\frac{1}{2}, m+1)} \left( 1 - \frac{x^2}{a^2} \right)^m, -a \leq x \leq a$$

$$(iv) f(x) = \frac{1}{a B(\frac{1}{2}, m-1)} \left( 1 + \frac{x^2}{a^2} \right)^m, -a \leq x \leq a,$$

8. Show that the Pearsonian Type VI curve may be written

$$y = y_0 \left( 1 - \frac{x^2}{a^2} \right)^{-\frac{m}{2}} \exp \left\{ -v \tanh^{-1} \frac{x}{a} \right\}$$

and discuss its relationship with Type IV curve.

9. Show that for Pearson distribution :

$$\frac{d}{dx} (\log f) = \frac{x}{B_0 + B_1 x + B_2 x^2}$$

the range is unlimited in both the directions if  $B_0 + B_1 x + B_2 x^2$  has no real roots, limited in one direction if roots are real and of the same sign, and limited in both directions if the roots are real and of opposite sign.

10. Investigate the properties and shapes which may be assumed by the frequency curve  $y = f(x)$  which has the differential equation

$$\frac{1}{y} \cdot \frac{dy}{dx} = - \frac{2mx}{a^2 - x^2}$$

and obtain the probability integral.

$$\text{Hint. } \frac{d}{dx} (\log y) = - \frac{2mx}{a^2 - x^2} \Rightarrow \log y = m \log (a^2 - x^2) + \log C$$

$$y = k \left( 1 - \frac{x^2}{a^2} \right)^m ; -a \leq x \leq a$$

which is type II distribution.

11. A family of distributions is defined by

$$\frac{1}{f} \cdot \frac{df}{dx} = \frac{x}{b_0 + b_2 x^2 + b_4 x^4}$$

and the frequency function  $f = f(x)$  vanishes at the terminals of its range. Show that the moments about the mean are given by

$$b_0(2s+1)\mu_{2s+1} + b_2(2s+3)\mu_{2s+3} + b_4(2s+5)\mu_{2s+5} = -\mu_{2s+2}$$

8.13. Variate Transformations. Let  $T$  be any statistic which is asymptotically normally distributed with mean  $\theta$  and variance  $\psi(\theta)$ , where  $\psi(\theta)$  is some function of the parameter  $\theta$  i.e.,  $T \sim N(\theta, \psi(\theta))$ . Let us transform  $T$  by a function  $g$  as  $g(T)$ , where  $g$  is a function which possesses first order derivative which is continuous and  $g'(\theta) \neq 0$ , where  $(')$  denotes differentiation w.r.t. the parameter  $\theta$ . Then  $g(T)$  is normally distributed about mean  $g(\theta)$  and variance  $[g'(\theta)]^2 \cdot \psi(\theta)$ , i.e.,

$$g(T) \sim N [ g(\theta), \{ g'(\theta) \}^2 \psi(\theta) ], \quad \dots(8.47)$$

asymptotically, provided  $g'(\theta) \neq 0$  is continuous in the neighbourhood of  $\theta$ .

In general  $\text{Var}(T) = \psi(\theta)$  will be dependent on the parameter  $\theta$ . We are interested in obtaining a function  $g$  such that the asymptotic variance of the transformed statistic  $g(T)$  is independent of  $\theta$ , i.e., it should be constant. In other words, we want  $g$  such that

$$\text{Var}[g(T)] = [g'(\theta)]^2 \cdot \psi(\theta) = \text{constant} = c^2, \text{ (say)}$$

$$\text{i.e., } [g'(\theta)] = \frac{c}{\sqrt{\psi(\theta)}}$$

Integrating both sides w.r.t.  $\theta$ , we get

$$g(\theta) = \int \frac{c}{\sqrt{\psi(\theta)}} d\theta \quad \dots(8.48)$$

8.13.1. Uses of Variate Transformations. As discussed above, when we transform a statistic  $T$  to a function  $g(T)$ , the distribution of  $g(T)$  is approximately

normal and its asymptotic variance is independent of the population parameter  $\theta$ . Hence the use of statistic  $g(T)$  gives better results and confidence intervals than the original statistic  $T$ . The commonly used transformations are :

- (1) Square Root Transformation.
- (2) Sine Inverse or  $\sin^{-1}$  Transformation.
- (3) Logarithmic Transformation.
- (4) Fisher's Z-Transformation.

In the following sections we shall discuss these transformations briefly.

**8-13-2. Square Root Transformation.** Square root transformation is a transformation for the Poisson variate. If a variable  $X$  follows Poisson distribution with parameter  $\lambda$  (assumed to be large), then we know that asymptotic distribution of  $X$  is normal as ( $\lambda \rightarrow \infty$ ) with  $E(X) = \lambda$  and  $\text{Var}(X) = \lambda = \psi(\lambda)$ , in the above notations. Then (8-48) gives the function

$$g(\lambda) = \int \frac{c}{\sqrt{\lambda}} d\lambda = 2c\sqrt{\lambda}.$$

Now we select  $c$  in such a way that  $2c = 1$

i.e.,  $c = 1/2$  so that  $g(\lambda) = \sqrt{\lambda}$ .

Hence the transformed variable is  $g(\bar{X}) = \sqrt{\bar{X}}$ . Using (8-47), the transformed variable  $\sqrt{\bar{X}}$  has the mean

$$\begin{aligned} g(\lambda) &= \sqrt{\lambda} \\ \text{and } \text{Var}(\sqrt{\bar{X}}) &= [g'(\lambda)]^2 \cdot \psi(\lambda) \\ &= \left( \frac{1}{2\sqrt{\lambda}} \right)^2 \cdot \lambda \\ &= 1/4. \end{aligned}$$

or alternatively

$$\text{Var}(\sqrt{\bar{X}}) = \text{constant} = c^2 = (1/2)^2 = 1/4.$$

Hence  $\sqrt{\bar{X}} \sim N(\sqrt{\lambda}, 1/4)$ , asymptotically. ... (8-49)

Ancombi has suggested the transformation  $\sqrt{\bar{X} + b}$ , where  $b$  is a constant suitably chosen.

**8-13-3. Sine Inverse or  $\sin^{-1}$  Transformation.** Sine-inverse is the transformation for stabilizing the variance of a binomial variate. If  $p$  is the observed proportion of successes in a series of  $n$  independent trials with constant probability  $P$  of success for each trial then we know that the asymptotic distribution of  $p$  is asymptotically normal (as  $n \rightarrow \infty$ ) with  $E(p) = P$  and

$$\text{Var}(p) = PQ/n, \quad Q = 1 - P$$

$$\text{i.e., } p \sim N\left(P, \frac{PQ}{n}\right), \text{ as } n \rightarrow \infty.$$

In the usual notations

$$\psi(P) = \frac{PQ}{n} = \frac{P(1-P)}{n}.$$

Using (8-48), the transforming function

$$\begin{aligned}
 g(P) &= \int \frac{c}{\sqrt{\psi(P)}} dP = c \sqrt{n} \int \frac{dP}{\sqrt{P}(1-P)} \\
 &= 2c \sqrt{n} \sin^{-1}(\sqrt{P}) \\
 &\quad \left[ \because \frac{d}{dP} \sin^{-1}(\sqrt{P}) = \frac{1}{2\sqrt{P}\sqrt{1-P}} \right]
 \end{aligned}$$

Choosing the constant  $c$  so that

$$2c \sqrt{n} = 1 \Rightarrow c = \frac{1}{2\sqrt{n}},$$

we get

$$g(P) = \sin^{-1}(\sqrt{P})$$

Hence the transformed statistic is  $g(p) = \sin^{-1}\sqrt{p}$ . Using (8.47),  $g(p)$  has mean  $\sin^{-1}\sqrt{P}$  and

$$\begin{aligned}
 \text{Var}(\sin^{-1}\sqrt{P}) &= [g'(P)]^2 \cdot \psi(P) \\
 &= \left[ \frac{1}{2\sqrt{P}\sqrt{1-P}} \right]^2 \times \frac{P(1-P)}{n} \\
 &= \frac{1}{4n}
 \end{aligned}$$

$$\text{or } \text{Var}(\sin^{-1}\sqrt{p}) = c^2 = \left( \frac{1}{2\sqrt{n}} \right)^2 = \frac{1}{4n} = \text{constant.}$$

$$\text{Hence } \sin^{-1}\sqrt{p} \sim N\left(\sin^{-1}\sqrt{P}, \frac{1}{4n}\right), \text{ asymptotically.} \quad \dots(8.50)$$

If  $r$  is the observed number of successes in  $n$  trials so that  $p = r/n$ , then Ancombi has suggested that instead of  $\sin^{-1}\sqrt{p} = \sin^{-1}\sqrt{r/n}$ , the transformation should be  $\sin^{-1}\sqrt{\frac{r+3/8}{n+3/8}}$ .

**8.13.4. Logarithmic Transformation.** Log transformation is the transformation for stabilizing the variance of the distribution of sample variance. If  $s^2$  is the sample variance in a sample of size  $n$  from normal population with variance  $\sigma^2$ , then the sampling distribution of  $s^2$  is asymptotically normal (as  $n \rightarrow \infty$ ) with

$$E(s^2) = \sigma^2 \text{ and } \text{var}(s^2) = \frac{2\sigma^4}{n} \text{ (for large } n\text{),}$$

[c.f. Remark to Theorem 13.5]. In the usual notations we have.

$\psi(\sigma^2) = \frac{2}{\sqrt{2}} \sigma^2/n$ . Using (8.48), the transforming function is

$$g(\sigma^2) = \int \frac{c\sqrt{n}}{\sqrt{2}\sigma^2} d\sigma^2 = \frac{c\sqrt{n}}{\sqrt{2}} \log \sigma^2.$$

We select  $c$  in such a way that

$$\frac{c\sqrt{n}}{\sqrt{2}} = 1 \Rightarrow c = \frac{\sqrt{2}}{\sqrt{n}},$$

so that  $g(\sigma^2) = \log_e \sigma^2$ .

Hence the transformation for the statistic  $s^2$  is  $g(s^2) = \log s^2$  and using (8.47) the transformed statistic is normally distributed with mean

$$g(\sigma^2) = \log \sigma^2 \text{ and}$$

$$\begin{aligned}\text{Var}[g(s^2)] &= [g'(\sigma^2)]^2 \cdot \psi(\sigma^2) \\ &= \left(\frac{1}{\sigma^2}\right)^2 \cdot \frac{2\sigma^4}{n} \\ &= \frac{2}{n}\end{aligned}$$

$$\text{or} \quad \text{Var}[g(s^2)] = c^2 = \left(\frac{\sqrt{2}}{\sqrt{n}}\right)^2 = \frac{2}{n}.$$

$$\text{Hence} \quad \log_e s^2 \sim N\left(\log_e \sigma^2, \frac{2}{n}\right), \text{ for large } n. \quad \dots(8.51)$$

**8.13.5. Fisher's z-Transformation.** This transformation is suggested for stabilizing the variance of sampling distribution of correlation coefficient (c.f. chapter 11). If  $r$  is the sample correlation coefficient in sampling from a correlated bivariate normal population with correlation coefficient  $\rho$  then the asymptotic distribution of  $r$ , as  $n \rightarrow \infty$  is normal with  $E(r) = \rho$  and

$$\text{Var}(r) = \frac{(1 - \rho^2)^2}{n} = \psi(\rho) \text{ for large } n. \text{ Using (8.47), we get}$$

$$g(\rho) = \int \frac{\sqrt{n} c}{1 - \rho^2} d\rho = \frac{\sqrt{n} c}{2} \log_e \left( \frac{1 + \rho}{1 - \rho} \right)$$

We select  $c$  in such a way that

$$\sqrt{n} c = 1 \Rightarrow c = 1/\sqrt{n},$$

$$\text{so that} \quad g(\rho) = \frac{1}{2} \log_e \left( \frac{1 + \rho}{1 - \rho} \right).$$

Hence using (8.47), the transformed statistic

$$g(r) = \frac{1}{2} \log_e \left( \frac{1 + r}{1 - r} \right), \text{ which is denoted by } Z, \text{ is normally distributed with mean}$$

$$g(\rho) = \frac{1}{2} \log_e \left( \frac{1 + \rho}{1 - \rho} \right) \text{ and}$$

$$\text{Var}[g(r)] = c^2 = \frac{1}{n}$$

$$\text{or} \quad \text{Var}[g(r)] = [g'(\rho)]^2 \psi(\rho)$$

$$= \left[ \frac{1}{1 - \rho^2} \right]^2 \cdot \frac{(1 - \rho^2)^2}{n}$$

$$= \frac{1}{n}, \text{ for large } n$$

Hence  $Z = \frac{1}{2} \log_e \left[ \frac{1+r}{1-r} \right] \sim N \left[ \frac{1}{2} \log_e \frac{1+\rho}{1-\rho}, \frac{1}{n} \right], \text{ for large } n.$  ... (8.52)

Prof. R.A. Fisher proved that the transformed statistic  $Z = g(r)$  is normally distributed even if  $n$  is small and that for exact samples (small  $n$ ),

$$Z \sim N \left[ \frac{1}{2} \log_e \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right].$$

For the various applications of this transformation the reader is referred to § 14.7.2.

**Remark.** We have :  $Z = \frac{1}{2} \log_e \left[ \frac{1+r}{1-r} \right] = \tanh^{-1}(r)$

Hence Z-transformation is also called the tan-hyperbolic-inverse transformation.

**8.14 Order Statistics** Let  $X_1, X_2, \dots, X_n$  be  $n$  independent and identically distributed variates, each with cumulative distribution function  $F(x)$ . If these variables are arranged in ascending order of magnitude and then written as  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ , we call  $X_{(r)}$  as the  $r$ th order statistic,  $r = 1, 2, \dots, n$ . The  $X_{(r)}$ 's because of the inequality relations among them are necessarily dependent.

**Remark.** If we write these ordered values as

$Y_1 \leq Y_2 \leq \dots \leq Y_n$ , then:

$Y_r = X_{(r)} = r$ th smallest of  $X_1, X_2, \dots, X_n$

$Y_1 = X_{(1)} =$  The smallest of  $X_1, X_2, \dots, X_n$

$Y_n = X_{(n)} =$  The largest of  $X_1, X_2, \dots, X_n$

#### 8.14.1. Cumulative Distribution Function of a Single Order Statistic.

Let  $F_r(x)$ ,  $r = 1, 2, \dots, n$  denote the c.d.f. of the  $r$ th order statistic  $X_{(r)}$ . Then the c.d.f. of the largest order statistic  $X_{(n)}$  is given by :

$$\begin{aligned} F_n(x) &= P(X_{(n)} \leq x) = P(X_i \leq x ; i = 1, 2, \dots, n) \\ &= P(X_1 \leq x \cap X_2 \leq x \cap \dots \cap X_n \leq x) \\ &= P(X_1 \leq x) \cdot P(X_2 \leq x) \dots P(X_n \leq x) \quad (\because X_i \text{'s are independent}) \\ &= [F(x)]^n, \end{aligned} \quad \dots (8.53)$$

since  $X_1, X_2, \dots, X_n$  are identically distributed.

The c.d.f. of the smallest order statistic  $X_{(1)}$  is given by :

$$\begin{aligned} F_1(x) &= P(X_{(1)} \leq x) \\ &= 1 - P(X_{(1)} > x) \\ &= 1 - P[X_i > x ; i = 1, 2, \dots, n] \\ &= 1 - \prod_{i=1}^n P(X_i > x) = 1 - \prod_{i=1}^n [1 - P(X_i \leq x)] \\ &= 1 - [1 - F(x)]^n, \end{aligned} \quad \dots (8.54)$$

since  $X_1, X_2, \dots, X_n$  are i.i.d. rv's.

In general, the c.d.f. of the  $r$ th order statistic  $X_{(r)}$  is given by:

$$\begin{aligned} F_r(x) &= P(X_{(r)} \leq x) \\ &= P[\text{At least } r \text{ of the } X_i \text{'s are } \leq x] \\ &= \sum_{j=r}^n P[\text{Exactly } j \text{ of the } n, X_i \text{'s are } \leq x], \\ &= \sum_{j=r}^n \binom{n}{j} F^j(x) [1 - F(x)]^{n-j}, \end{aligned} \quad \dots(8.55)$$

by using Binomial probability model.

**Remarks.** 1. (8.55) can also be written as [See Remark 2 to Example 7.23]:

$$F_r(x) = I_{F(x)}(r, n - r + 1), \quad \dots(8.56)$$

$$\text{where } I_p(a, b) = \frac{1}{\beta(a, b)} \int_0^p t^{a-1} (1-t)^{b-1} dt \quad \dots(8.56a)$$

is the 'incomplete Beta Function' and has been tabulated in Biometrika tables by Pearson and Hartley.

(8.56) and (8.56a) show that the probability points of an order statistic can be obtained with the help of incomplete beta function.

2. Taking  $r = 1$  and  $r = n$  in (8.55), we get respectively:

$$\begin{aligned} F_1(x) &= \sum_{j=1}^n \binom{n}{j} F^j(x) [1 - F(x)]^{n-j} \\ &= 1 - \left\{ \left[ \binom{n}{j} F^j(x) [1 - F(x)]^{n-j} \right] \right\}_{j=0} \\ &= 1 - [1 - F(x)]^n \end{aligned} \quad \dots(8.56b)$$

$$\text{and } F_n(x) = F^n(x), \quad \dots(8.56c)$$

the results which have already been obtained in (8.54) and (8.53) respectively,

**8.14.2. Probability Density Function (p.d.f.) of a Single Order Statistic.** The results in (8.53) to (8.55) are valid for both discrete and continuous r.v.'s. We shall now assume that  $X_i$ 's are i.i.d. continuous r.v.'s with p.d.f.  $f(x) = F'(x)$ . If  $f_r(x)$  denotes the p.d.f. of  $X_{(r)}$  then from (8.55) or (8.56) we get:

$$\begin{aligned} f_r(x) &= \frac{d}{dx} [F_r(x)] = \frac{d}{dx} [I_{F(x)}(r, n - r + 1)] \\ &= \frac{d}{dx} \left[ \frac{1}{\beta(r, n - r + 1)} \int_0^{F(x)} t^{r-1} (1-t)^{n-r} dt \right] \end{aligned} \quad \dots(8.57)$$

Let us write

$$g(t) = \int t^{r-1} (1-t)^{n-r} dt \Rightarrow g'(t) = t^{r-1} (1-t)^{n-r} \quad \dots(*)$$

$$\Rightarrow \int_0^{F(x)} t^{r-1} (1-t)^{n-r} dt = [g(t)]_0^{F(x)} = g(F(x)) - g(0)$$

$$\Rightarrow \frac{d}{dx} \int_0^{F(x)} t^{r-1} (1-t)^{n-r} dt = g'(F(x)) \cdot f(x) \quad (\because g(0) \text{ is constant})$$

$$= [F(x)]^{r-1} [1-F(x)]^{n-r} f(x) \quad [\text{Using } (*)]$$

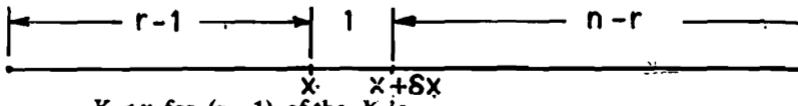
Substituting in (8.57) we get :

$$f_r(x) = \frac{1}{\beta(r, n-r+1)} \cdot F'^{-1}(x) [1-F(x)]^{n-r} \cdot f(x) \quad ... (8.58)$$

Aliter. By definition of a p.d.f. we get:

$$f_r(x) = \lim_{\delta x \rightarrow 0} \frac{P[x < X_{(r)} \leq x + \delta x]}{\delta x} \quad ... (8.59)$$

The event  $E: x < X_{(r)} \leq x + \delta x$  can materialise as follows:



$X_i \leq x$  for  $(r-1)$  of the  $X_i$ 's

$x < X_i \leq x + \delta x$  for one  $X_i$

and  $X_i \geq x + \delta x$  for the remaining  $(n-r)$  of the  $X_i$ 's.

Hence by the multinomial probability law we have:

$$P(x < X_{(r)} \leq x + \delta x) = \frac{n!}{(r-1)! 1! (n-r)!} p_1^{r-1} \cdot p_2^1 \cdot p_3^{n-r} \quad ... (8.60)$$

where  $p_1 = P(X_i \leq x) = F(x)$

$p_2 = P(x < X_i \leq x + \delta x) = F(x + \delta x) - F(x)$

and  $p_3 = P(X_i \geq x + \delta x) = 1 - P(X_i \leq x + \delta x) = 1 - F(x + \delta x)$

Substituting in (8.60), we get:

$$f_r(x) = \lim_{\delta x \rightarrow 0} \frac{P(x < X_{(r)} \leq x + \delta x)}{\delta x}$$

$$= \frac{1}{\beta(r, n-r+1)} \times F'^{-1}(x) \times \lim_{\delta x \rightarrow 0} \left[ \frac{F(x + \delta x) - F(x)}{\delta x} \right]$$

$$\times \lim_{\delta x \rightarrow 0} [1 - F(x + \delta x)]^{n-r}$$

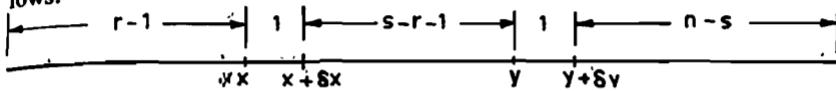
$$= \frac{1}{\beta(r, n-r+1)} \cdot F'^{-1}(x) \cdot f(x) \cdot [1 - F(x)]^{n-r},$$

as in (8.58).

**8.14.3. Joint p.d.f. of two Order Statistics.** Let us denote the joint p.d.f. of  $X_{(r)}$  and  $X_{(s)}$ , where  $1 \leq r < s \leq n$  by  $f_{rs}(x, y)$ . Then,

$$f_{rs}(x, y) = \lim_{\delta x \rightarrow 0} \lim_{\delta y \rightarrow 0} \frac{P[x \leq X_{(r)} \leq x + \delta x \cap y \leq X_{(s)} \leq y + \delta y]}{\delta x \delta y} \quad ... (8.6)$$

The event  $E = \{x \leq X_{(r)} \leq x + \delta x \cap y \leq X_{(s)} \leq y + \delta y\}$  can materialise as follows:



$X_i \leq x$  for  $r-1$  of the  $X_i$ 's,

$x < X_i \leq x + \delta x$  for one  $X_i$ ,

$x + \delta x < X_i \leq y$  for  $(s-r-1)$  of  $X_i$ 's,

$y < X_i \leq y + \delta y$  for one  $X_i$ .

and  $X_i > y + \delta y$  for  $(n-s)$  of the  $X_i$ 's

Hence using multinomial probability law, we get

$$\begin{aligned} P(E) &= P\left[\frac{x \leq X_{(r)} \leq x + \delta x \cap y \leq X_{(s)} \leq y + \delta y}{n!}\right] \\ &= \frac{n!}{(r-1)! 1! (s-r-1)! 1! (n-s)!} p_1^{r-1} \cdot p_2 p_3^{s-r-1} \cdot p_4 p_5^{n-s} \end{aligned} \quad \dots(8-62)$$

where

$$p_1 = P(X_i \leq x) = F(x)$$

$$p_2 = P(x < X_i \leq x + \delta x) = F(x + \delta x) - F(x)$$

$$p_3 = P(x + \delta x < X_i \leq y) = F(y) - F(x + \delta x)$$

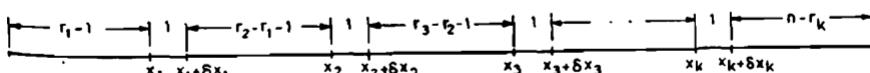
$$p_4 = P(y < X_i \leq y + \delta y) = F(y + \delta y) - F(y)$$

$$p_5 = P(X_i > y + \delta y) = 1 - P(X_i \leq y + \delta y) = 1 - F(y + \delta y)$$

Substituting in (8-62) and using (8-61) we get:

$$\begin{aligned} f_{rs}(x, y) &= \lim_{\delta x \rightarrow 0} \lim_{\delta y \rightarrow 0} \frac{P(E)}{\delta x \delta y} \\ &= \frac{n!}{(r-1)! (s-r-1)! (n-s)!} \times F^{r-1}(x) \times \lim_{\delta x \rightarrow 0} \frac{[F(x + \delta x) - F(x)]}{\delta x} \\ &\quad \times \lim_{\delta y \rightarrow 0} \left[ \frac{F(y + \delta y) - F(y)}{\delta y} \right] \times \lim_{\delta y \rightarrow 0} [1 - F(y + \delta y)]^{n-s} \\ &\quad \times \lim_{\delta x \rightarrow 0} [F(y) - F(x + \delta x)]^{s-r-1} \\ &= \frac{n!}{(r-1)! (s-r-1)! (n-s)!} F^{r-1}(x) \cdot f(x) \cdot [F(y) - F(x)]^{s-r-1} f(y) \cdot [1 - F(y)]^{n-s} \end{aligned} \quad \dots(8-63)$$

**8-14-4 Joint p.d.f. of  $k$ -Order Statistics.** The joint p.d.f. of  $k$ -order statistics  $X_{(r_1)}, X_{(r_2)}, \dots, X_{(r_k)}$  where  $1 \leq r_1 < r_2 < \dots < r_k \leq n$  and  $1 \leq k \leq n$  is for  $x_1 \leq x_2 \leq \dots \leq x_k$  given by [on using the following configuration and the multinomial probability law as in § 8-14-3]:

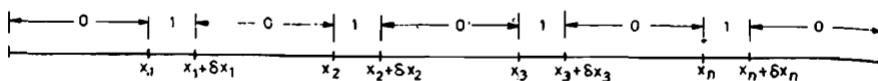


$$f_{r_1, r_2, \dots, r_k}(x_1, x_2, \dots, x_k) = \frac{n!}{(r_1-1)! (r_2-r_1-1)! \dots (r_k-r_{k-1}-1)! (n-r_k)!} \\ \times F'^{r_1-1}(x_1) \times f(x_1) \times \left[ F(x_2) - F(x_1) \right]^{r_2-r_1-1} \times f(x_2) \\ \times \left[ F(x_3) - F(x_2) \right]^{r_3-r_2-1} \times f(x_3) \times \dots \times f(x_k) \left[ 1 - F(x_k) \right]^{n-r_k} \quad \dots(8.64)$$

**8.14-5. Joint p.d.f. of all  $n$  - Order Statistics.** In particular the joint p.d.f. of all the  $n$  order statistics is obtained on taking  $k = n$  in (8.64). This implies that  $r_i = i$  for  $i = 1, 2, \dots, n$ . Hence joint.p.d.f. of  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  is given by:

$$f_{1, 2, \dots, n}(x_1, x_2, \dots, x_n) = n! f(x_1) f(x_2) \dots f(x_n) \quad \dots(8.65)$$

**Aliter.** We can easily obtain (8.65) by using the following configuration:



and the multinomial probability law as in § 8.14-3.

**8.14-6. Distribution of Range and other Systematic Statistics.** Let us obtain the p.d.f. the statistic  $W_{rs} = X_{(s)} - X_{(r)}$ ;  $r < s$ . We start with the joint p.d.f. of  $X_{(r)}$  and  $X_{(s)}$  given in (8.63) and transform  $[X_{(r)}, X_{(s)}]$  to the new variables  $W_{rs}$  and  $\bar{x}$  s.t.

$$w_{rs} = y - x; \quad x = \bar{x}; \quad \text{s.t.}; \quad y = x + w_{rs} \quad \text{and} \quad \bar{x} = x$$

$$\therefore J = \begin{vmatrix} \frac{\partial(x, y)}{\partial(\bar{x}, w_{rs})} & \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} \end{vmatrix} = 1 \Rightarrow |J| = 1$$

The joint p.d.f.  $f_{rs}(x, y)$  in (8.63) transforms to the joint p.d.f. of  $X_{(r)}$  and  $W_{rs}$  as given below:

$$g(\bar{x}, w_{rs}) = c_{rs} \cdot F'^{r-1}(x) \cdot f(x) \left[ F(x + w_{rs}) - F(x) \right]^{s-r-1} \\ \times f(x + w_{rs}) \times \left[ 1 - F(x + w_{rs}) \right]^{n-s} \quad \dots(8.66)$$

$$\text{where } c_{rs} = \frac{n!}{(r-1)! (s-r-1)! (n-s)!} \quad \dots(8.67)$$

Integrating (8.66) w.r.to.  $x$  from  $-\infty$  to  $\infty$ , we obtain the p.d.f. of  $W_{rs}$  as:

$$g(w_{rs}) = c_{rs} \int_{-\infty}^{\infty} \left\{ F'^{r-1}(x) f(x) \left[ F(x + w_{rs}) - F(x) \right]^{s-r-1} \cdot f(x + w_{rs}) \right. \\ \left. \cdot \left[ 1 - F(x + w_{rs}) \right]^{n-s} \right\} dx \quad \dots(8.68)$$

**Remark. Distribution of Range  $W = X_{(n)} - X_{(1)}$ .** Taking  $r = 1$  and  $s = n$  in (8.68), we obtain the p.d.f. of the range  $W = X_{(n)} - X_{(1)}$  as:

$$g(w) = n(n-1) \int_{-\infty}^{\infty} f(x) \left[ F(x + w) - F(x) \right]^{n-2} \cdot f(x + w) dx; w \geq 0 \quad \dots(8.69)$$

The c.d.f of  $W$  is rather simple as given below:

$$\begin{aligned}
 G(w) &= P(W \leq w) = \int_0^w g(u) du \\
 &= \int_0^w \left\{ n(n-1) \int_{-\infty}^{\infty} f(x) [F(x+u) - F(x)]^{n-2} f(x+u) dx \right\} du \\
 &= n \int_{-\infty}^{\infty} f(x) \left\{ \int_0^w (n-1) f(x+u) (F(x+u) - F(x))^{n-2} du \right\} dx \\
 &= n \int_{-\infty}^{\infty} f(x) [F(x+w) - F(x)]^{n-1} dx
 \end{aligned}$$

**Example 8.48** Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with continuous density. Show that  $Y_1 = \min(X_1, X_2, \dots, X_n)$ , is exponential with parameter  $n\lambda$  if and only if each  $X_i$  is exponential with parameter  $\lambda$ .

**Solution.** Let  $X_i$  be i.i.d. exponential variates with parameter  $\lambda$  and p.d.f.

$$f(x) = \lambda e^{-\lambda x}; x \geq 0, \lambda > 0 \quad \dots(i)$$

$$F(x) = P(X \leq x) = \int_0^x f(u) du = \lambda \int_0^x e^{-\lambda u} du = 1 - e^{-\lambda x} \quad \dots(ii)$$

Distribution function  $G(\cdot)$  of  $Y_1 = \min(X_1, X_2, \dots, X_n)$  is given by:

$$G_{Y_1}(y) = P(Y_1 \leq y) = 1 - [1 - F(y)]^n \quad [\text{From (8.54)}]$$

$$= 1 - [1 - (1 - e^{-\lambda y})]^n = 1 - e^{-n\lambda y} \quad \dots(iii)$$

[From (ii)]

which is the distribution function of exponential distribution with parameter  $n\lambda$ . Hence  $Y_1 = \min(X_1, X_2, \dots, X_n)$ , has exponential distribution with parameter  $n\lambda$ .

Conversely, Let  $Y_1 = \min(X_1, X_2, \dots, X_n) \sim \text{Exp}(n\lambda)$  so that

$$P(Y_1 \leq y) = 1 - e^{-n\lambda y} \Rightarrow P(Y_1 \geq y) = e^{-n\lambda y} \quad \dots(iv)$$

$$\Rightarrow P[\min(X_1, X_2, \dots, X_n) \geq y] = e^{-n\lambda y}$$

$$\Rightarrow P[(X_1 \geq y) \cap (X_2 \geq y) \cap \dots \cap (X_n \geq y)] = e^{-n\lambda y}$$

$$\Rightarrow \prod_{i=1}^n P(X_i \geq y) = e^{-n\lambda y}$$

$$[P(X_i \geq y)]^n = e^{-n\lambda y} \quad \{ \because X_i's \text{ are i.i.d.} \}$$

$$\Rightarrow P(X_i \geq y) = e^{-\lambda y}$$

$$\Rightarrow P(X_i \leq y) = 1 - e^{-\lambda y}$$

which is the distribution function of  $\text{Exp}(\lambda)$  distribution. Hence  $X_i$ 's are i.i.d.  $\text{Exp}(\lambda)$ .

**Example 8-49.** For the exponential distribution  $f(x) = e^{-x}, x \geq 0$ ; show that the cumulative distribution function (c.d.f.) of  $X_{(n)}$  in a random sample of size  $n$  is  $F_n(x) = (1 - e^{-x})^n$ . Hence prove that as  $n \rightarrow \infty$ , the c.d.f. of  $X_{(n)} - \log n$  tends to the limiting form  $\exp[-(\exp(-x))]$ ,  $-\infty < x < \infty$ .

**Solution.** Here  $f(x) = e^{-x}, x \geq 0$ ;  $F(x) = P(X \leq x) = 1 - e^{-x}$  ...(\*)

The c.d.f. of  $X_{(n)}$  is given by [From (8-53)]

$$F_n(x) = P[X_{(n)} \leq x] = [F(x)]^n = (1 - e^{-x})^n \quad [\text{From } (*)] \dots (**)$$

The c.d.f.  $G_n(\cdot)$  of  $X_{(n)} - \log n$  is given by:

$$\begin{aligned} G_n(x) &= P[X_{(n)} - \log n \leq x] \\ &= P[X_{(n)} \leq x + \log n] \\ &= \left[ 1 - e^{-(x + \log n)} \right]^n \quad [\text{From } (**)] \\ &= \left[ 1 - \frac{e^{-x}}{n} \right]^n \quad \left[ \because e^{-\log n} = e^{\log n^{-1}} = \frac{1}{n} \right] \end{aligned}$$

$$\therefore \lim_{n \rightarrow \infty} G_n(x) = \lim_{n \rightarrow \infty} \left[ 1 - \frac{e^{-x}}{n} \right]^n = \exp[-e^{-x}] \quad \left[ \because \lim_{n \rightarrow \infty} \left( 1 + \frac{m}{n} \right)^n = e^m \right]$$

**Example 8-50** Show that for a random sample of size 2 from  $N(0, \sigma^2)$  population,  $E(X_{(1)}) = -\sigma/\sqrt{\pi}$  [Delhi Univ. M.Sc. (Stat.), 1988, 1982]

**Solution.** For  $n = 2$ , the p.d.f.  $f_1(x)$  of  $X_{(1)}$  is given by: [From (8-58)]

$$f_1(x) = \frac{1}{\beta(1, 2)} [1 - F(x)] f(x) = 2 [1 - F(x)] \cdot f(x); -\infty < x < \infty$$

$$\text{where } f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} \quad [\because X \sim N(0, \sigma^2)]$$

$$\therefore E(X_{(1)}) = \int_{-\infty}^{\infty} x \cdot f_1(x) dx = 2 \int_{-\infty}^{\infty} [1 - F(x)] \cdot x f(x) dx \quad \dots(i)$$

$$\text{We have: } \log f(x) = -\log(\sqrt{2\pi}\sigma) - \frac{x^2}{2\sigma^2}$$

Differentiating w.r.t.  $x$  we get:

$$\frac{f'(x)}{f(x)} = -\frac{x}{\sigma^2}$$

$$\Rightarrow \int x f(x) dx = -\sigma^2 \int f'(x) dx = -\sigma^2 f(x) \quad \dots(ii)$$

Integrating (i) by parts and using (ii), we get:

$$\begin{aligned}
 E(X_{(1)}) &= 2 \cdot \left[ [1 - F(x)] (-\sigma^2 f(x)) \right]_{-\infty}^{\infty} - 2 \int_{-\infty}^{\infty} (-\sigma^2 f(x)) (-f'(x)) dx \\
 &= -2 \sigma^2 \int_{-\infty}^{\infty} [f(x)]^2 dx = -\frac{1}{\pi} \int_{-\infty}^{\infty} e^{-x^2/\sigma^2} dx \\
 &= -\frac{1}{\pi} \cdot \frac{\sqrt{\pi}}{(1/\sigma)} \\
 &= -\sigma/\sqrt{\pi} \quad , \quad \left( \because \int_{-\infty}^{\infty} e^{-a^2 x^2} dx = \sqrt{\pi}/a \right)
 \end{aligned}$$

**Example 8-51.** Show that in odd samples of size  $n$  from  $U[0, 1]$  population, the mean and variance of the distribution of median are  $1/2$  and  $1/[4(n+2)]$  respectively.

**Solution.** We have:  $f(x) = 1 ; 0 \leq x \leq 1$

$$F(x) = P(X \leq x) = \int_0^x f(u) du = \int_0^x 1 \cdot du = x$$

Let  $n = 2m + 1$  (odd), where  $m$  is a positive integer  $\geq 1$ . Then median observation is  $X_{(m+1)}$ . Taking  $r = (m+1)$  in (8.58), the p.d.f of median  $X_{(m+1)}$  is given by:

$$\begin{aligned}
 f_{m+1}(x) &= \frac{1}{\beta(m+1, m+1)} \cdot x^m (1-x)^m \\
 \therefore E(X_{(m+1)}) &= \frac{1}{\beta(m+1, m+1)} \cdot \int_0^1 x \cdot x^m (1-x)^m dx \\
 &= \frac{\beta(m+2, m+1)}{\beta(m+1, m+1)} \\
 &= \frac{\Gamma(m+2) \cdot \Gamma(m+1)}{\Gamma(m+3)} \times \frac{\Gamma(2m+2)}{\Gamma(m+1) \Gamma(m+1)} \\
 &= \frac{m+1}{2m+2} = \frac{1}{2} \quad (\text{On simplification})
 \end{aligned}$$

$$\begin{aligned}
 E(X_{(m+1)}^2) &= \int_0^1 x^2 f_{m+1}(x) dx = \frac{1}{\beta(m+1, m+1)} \cdot \int_0^1 x^{m+2} (1-x)^m dx \\
 &= \frac{\beta(m+3, m+1)}{\beta(m+1, m+1)} = \frac{m+2}{2(2m+3)}
 \end{aligned}$$

$$\begin{aligned}
 \therefore \text{Var}(X_{(m+1)}) &= E(X_{(m+1)}^2) - [E(X_{(m+1)})]^2 \\
 &= \frac{m+2}{2(2m+3)} - \frac{1}{4} = \frac{1}{4(2m+3)} = \frac{1}{4(n+2)}
 \end{aligned}$$

**Example 8-52.** Let  $X_1, X_2, \dots, X_n$  be i.i.d. non-negative random variables of the continuous type with p.d.f.  $f(\cdot)$  and distribution function  $F(\cdot)$ .

If  $E|X| < \infty$ , Show that  $E|X_{(r)}| < \infty$ .

(b) Write  $M_n = X_{(n)} = \max_{\infty} (X_1, X_2, \dots, X_n)$ . Show that

$$E(M_n) = E(M_{n-1}) + \int_0^{\infty} F^{n-1}(x) [1 - F(x)] dx ; n = 2, 3, \dots$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

Hence evaluate  $E(M_n)$  if  $X_1, X_2, \dots, X_n$  have common distribution function:

$$F(x) = x ; 0 < x < 1.$$

**Solution** (a)  $E|X_r| = \int_0^{\infty} |x| \cdot f_r(x) dx$

$$\begin{aligned} & (\because X \text{ is non-negative continuous r.v.}) \\ &= \int_0^{\infty} |x| \cdot \frac{n!}{(r-1)! (n-r)!} f(x) \cdot F^{r-1}(x) [1 - F(x)]^{n-r} dx \\ &\leq n \binom{n-1}{r-1} \cdot \int_0^{\infty} |x| f(x) dx \\ &\leq n \binom{n-1}{r-1} E|X| \end{aligned}$$

Hence  $E|X_r| < \infty$  if  $E|X| < \infty$ .

(b) The p.d.f.  $f_n(x)$ , of  $M_n = X_{(n)}$  is given by:

$$f_n(x) = n [F(x)]^{n-1} \cdot f(x)$$

$$E(M_n) = \lim_{a \rightarrow \infty} \int_0^a x f_n(x) dx \quad (\because X \geq 0, a.s.)$$

$$\therefore E(M_n) = \lim_{a \rightarrow \infty} n \int_0^a x [F(x)]^{n-1} \cdot f(x) dx$$

Integrating by parts we get:

$$\begin{aligned} E(M_n) &= n \lim_{a \rightarrow \infty} \left[ x \cdot \frac{F^n(x)}{n} \Big|_0^a - \int_0^a \frac{F^n(x)}{n} \cdot 1 dx \right] \\ &= \lim_{a \rightarrow \infty} \left[ a F^n(a) - \int_0^a F^n(x) dx \right] \\ &= \lim_{a \rightarrow \infty} \left[ \int_0^a (1 - F^n(x) - 1) dx + a F^n(a) \right] \end{aligned}$$

$$\begin{aligned}
 &= \lim_{a \rightarrow \infty} \left[ \int_0^a (1 - F^n(x)) dx - a + a F^n(a) \right] \\
 &= \lim_{a \rightarrow \infty} \left[ \int_0^a (1 - F^n(x)) dx - a (1 - F^n(a)) \right]
 \end{aligned}$$

Since  $E M_n$  exists, (By part (a)),

$$\begin{aligned}
 a \cdot P(M_n > a) &= a [1 - P(M_n \leq a)] = a [1 - F^n(a)] \\
 &\longrightarrow 0 \text{ as } a \rightarrow \infty.
 \end{aligned}$$

$$\begin{aligned}
 E(M_n) &= \lim_{a \rightarrow \infty} \int_0^a (1 - F^n(x)) dx = \int_0^\infty (1 - F^n(x)) dx \quad \dots(*) \\
 &= \int_0^\infty (1 - F^{n-1}(x) \cdot F(x)) dx \\
 &= \int_0^\infty [1 - F^{n-1}(x) [1 - (1 - F(x))] ] dx \\
 &= \int_0^\infty (1 - F^{n-1}(x)) dx + \int_0^\infty F^{n-1}(x) [1 - F(x)] dx \\
 &= E(M_{n-1}) + \int_0^\infty F^{n-1}(x) [1 - F(x)] dx \quad [\text{From } (*)] \dots(**)
 \end{aligned}$$

If  $X \sim U[0, 1]$ , then

$$f(x) = 1 ; \quad 0 < x < 1 \quad \text{and} \quad F(x) = x ; \quad 0 < x < 1$$

Substituting in (\*\*), we get:

$$\begin{aligned}
 E(M_n) - E(M_{n-1}) &= \int_0^1 x^{n-1} (1-x) dx \\
 \Rightarrow E(M_n) - E(M_{n-1}) &= \frac{1}{n} - \frac{1}{n+1} \quad \dots(***)
 \end{aligned}$$

Changing  $n$  to  $n-1, n-2, \dots, 2, 1$  in (\*\*\*) we get respectively:

$$E(M_{n-1}) - E(M_{n-2}) = \frac{1}{n-1} - \frac{1}{n}$$

$$E(M_2) - E(M_1) = \frac{1}{2} - 1$$

$$E(M_1) - E(M_0) = 1 - \frac{1}{2}$$

Adding (\*\*\* ) and the above equations and noting that  $E(M_0) = 0$ , we get;

$$E(M_n) = 1 - \frac{1}{n+1} = \frac{n}{n+1}$$

**Example 8.53 (a)** Find the p.d.f. of  $X_{(r)}$  in a random sample of size  $n$  from the exponential distribution:

$$f(x) = \alpha e^{-\alpha x}, \alpha > 0, x \geq 0$$

(b) Show that  $X_{(r)}$  and  $W_{rs} = X_{(s)} - X_{(r)}, r < s$ , are independently distributed.

(c) What is the distribution of  $W_1 = X_{(r+1)} - X_{(r)}$  ?

$$\text{Solution. Here } F(x) = P(X \leq x) = \int_0^x \alpha \cdot e^{-\alpha u} du = 1 - e^{-\alpha x} \quad \dots(i)$$

The p.d.f. of  $X_{(r)}$  is given by:

$$\begin{aligned} f_r(x) &= \frac{1}{\beta(r, n-r+1)} \cdot [F(x)]^{r-1} \cdot [1-F(x)]^{n-r} \cdot f(x) \\ &= \frac{1}{\beta(r, n-r+1)} \cdot (1 - e^{-\alpha x})^{r-1} \cdot e^{-\alpha x(n-r)} \cdot \alpha \cdot e^{-\alpha x} \\ &= \frac{1}{\beta(r, n-r+1)} \cdot \alpha \cdot e^{-\alpha x(n-r+1)} \cdot [1 - e^{-\alpha x}]^{r-1}; x > 0 \end{aligned}$$

(b) The joint p.d.f. of  $X_{(r)}$  and  $W_{rs} = X_{(s)} - X_{(r)}$  is given by [From (8.66)]

$$\begin{aligned} g(x, w_{rs}) &= c_{rs} \cdot F^{r-1}(x) f(x) \left[ F(x + w_{rs}) - F(x) \right]^{s-r-1} \\ &\quad \times f(x + w_{rs}) \left[ 1 - F(x + w_{rs}) \right]^{n-s} \\ &= \frac{n!}{(r-1)! (n-r)!} \times \frac{(n-r)!}{(s-r-1)! (n-s)!} \times [1 - e^{-\alpha x}]^{r-1} \alpha e^{-\alpha x} \\ &\quad \times \left[ e^{-\alpha x} - e^{-\alpha(x+w_{rs})} \right]^{s-r-1} \times \alpha e^{-\alpha(x+w_{rs})} \times \left[ e^{-\alpha(x+w_{rs})} \right]^{n-s} \\ &= \left[ \frac{1}{\beta(r, n-r+1)} \cdot \alpha e^{-\alpha x(n-r+1)} (1 - e^{-\alpha x})^{r-1} \right] \\ &\quad \times \left[ \frac{1}{\beta(s-r, n-s+1)} \cdot \alpha \cdot e^{-(n-s+1)\alpha w_{rs}} (1 - e^{\alpha w_{rs}})^{s-r-1} \right] \dots(ii) \end{aligned}$$

$\Rightarrow X_{(r)}$  and  $W_{rs}$  are independently distributed.

(c) Taking  $s = r+1$  in (ii), the p.d.f. of  $W_1 = X_{(r+1)} - X_{(r)}$  becomes:

$$\begin{aligned} g(w_1) &= \frac{1}{\beta(1, n-r)} \cdot \alpha \cdot e^{-\alpha(n-r)w_1} \\ &= (n-r) \alpha \cdot e^{-(n-r)\alpha w_1}; w_1 \geq 0 \end{aligned}$$

which shows that  $W_1$  has an exponential distribution with parameter  $(n-r)\alpha$ .

## EXERCISE 8 (j)

(1) (a) Obtain the distribution function and hence the p.d.f. of the  $r$ th order statistic  $X_{(r)}$  in a random sample of size  $n$  from a population with continuous distribution function  $P(\cdot)$ . Deduce the p.d.f.'s of the smallest and the largest sample observations. [Delhi Univ. M.Sc. (Stat.), 1987]

(b) Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a population having continuous distribution function  $P(x)$ . Define the  $r$ th order statistic  $X_{(r)}$  and obtain its distribution function and hence its p.d.f.

[Delhi Univ. M.Sc. (Stat.), 1983]

2. Define  $r$ th order statistic  $X_{(r)}$ . Obtain the joint p.d.f. of  $X_{(r)}$  and  $X_{(s)}$ ,  $r < s$ , in a random sample of size  $n$  from a population with continuous distribution function  $P(\cdot)$ . Hence deduce the p.d.f. of sample range  $W = X_{(n)} - X_{(1)}$ .

[Delhi Univ. M.Sc. (Stat.), 1988, 1982]

3. Obtain the distribution function and hence the p.d.f. of the smallest sample observation  $X_{(1)}$  in a random sample of size  $n$  from a population with a continuous distribution function  $F(x)$ . Show that for random sample of size 2 from normal population  $N(0, \sigma^2)$ ,  $E(X_{(1)}) = -\sigma/\sqrt{\pi}$

[Bombay Univ. M.Sc. (Stat.), 1992]

4. Let  $X_1, X_2, \dots, X_n$  be  $n$  independent variates,  $X_i$  having a geometric distribution with parameter  $p_i$ , i.e.,

$$P(X_i = x_i) = q_i^{x_i-1} \cdot p_i; \quad q_i = 1 - p_i, \quad x_i = 1, 2, 3, \dots$$

Show that  $X_{(1)}$  is distributed geometrically with parameter  $(1 - q_1 q_2 \dots q_n)$

[Delhi Univ. M.Sc. (Stat.), 1983]

(b) Let  $X_1, X_2, \dots, X_n$  be  $n$  independent variates,  $X_i$  having a geometric distribution with parameter  $p_i$  i.e.

$$P(X_i = x_i) = q_i^{x_i-1} \cdot p_i; \quad q_i = 1 - p_i, \quad x_i = 1, 2, 3, \dots,$$

Show that  $X_{(1)}$  is distributed geometrically with parameter  $(1 - q_1 q_2 q_3 \dots q_n)$ .

5. For a random sample of size  $n$  from a continuous population whose p.d.f.  $p(x)$  is symmetrical at  $x = \mu$ , show that

$$f_r(\mu + x) = f_{n-r+1}(\mu - x),$$

where  $f_r(\cdot)$  is the p.d.f. of  $X_{(r)}$ .

Hint.  $f(\mu + x) = f(\mu - x)$

$$\begin{aligned} F(\mu + x) &= P(X \leq \mu + x) = P(X \geq \mu - x) && \text{(By symmetry)} \\ &= 1 - P(X \leq \mu - x) = 1 - F(\mu - x). \end{aligned}$$

6. Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a population having continuous distribution function  $F(x)$ .

Define the order statistic of rank  $k$ ,  $1 \leq k \leq n$ . Find its distribution function.

Show that for the rectangular distribution

$$f(x) = 1/\theta_2, \quad \theta_1 - \frac{1}{2}\theta_2 \leq x \leq \theta_1 + \frac{1}{2}\theta_2,$$

$$E \left[ \frac{X_{(r)} - \theta_1}{\theta_2} \right] = \frac{r}{n+1} - \frac{1}{2}.$$

7. Let  $X_1, X_2, \dots, X_n$  be a random sample with common p.d.f.

$$f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(i) Find the p.d.f., mean and variance of  $X_{(1)}$ .

(ii) Find the p.d.f., mean and variance of  $X_{(n)}$ .

(iii) Find Corr. ( $X_{(1)}, X_{(n)}$ ).

**Ans.** (i)  $f_1(x) = n(1-x)^{n-1}, 0 \leq x \leq 1; E(X_{(1)}) = 1/(n+1)$

$$\text{Var}(X_{(1)}) = n/[ (n+2)(n+1)^2 ]$$

(ii)  $f_n(x) = nx^{n-1}; 0 \leq x \leq 1; E(X_{(n)}) = n/(n+1)$

$$\text{Var}(X_{(n)}) = n/[ (n+2)(n+1)^2 ]$$

(iii) **Hint.**  $r(X_{(1)}, X_{(n)}) = \frac{\text{Cov}(X_{(1)}, X_{(n)})}{\sqrt{\text{Var}(X_{(1)}) \cdot \text{Var}(X_{(n)})}}$  [c.f. Chapter 10]

$$E(X_{(1)} \cdot X_{(n)}) = \int_0^1 \int_0^y xy f_{1n}(x, y) dx dy = n(n-1) \int_0^1 \int_0^y xy(y-x)^{n-2} dx dy$$

$$\left( \because f_{1n}(x, y) = n(n-1)f(x)[F(y)-F(x)]^{n-2}f(y); 0 \leq x < y \leq 1 \right)$$

$$\therefore E(X_{(1)} \cdot X_{(n)}) = n(n-1) \int_0^1 \int_0^y x y^{n-1} \left(1 - \frac{x}{y}\right)^{n-2} dx dy$$

$$= n(n-1) \int_0^1 \int_0^t y^{n+1} t(1-t)^{n-2} dt dy; \left(\frac{x}{y} = t\right)$$

$$= 1/(n+2) \quad [\text{On simplification}]$$

$$\text{Cov}(X_{(1)}, X_{(n)}) = 1/[ (n+1)^2 \cdot (n+2) ]$$

$$\text{Corr.}(X_{(1)}, X_{(n)}) = 1/n$$

8. Show that the c.d.f. of the mid-point (or mid-range)  $M = \frac{1}{2}(X_{(1)} + X_{(n)})$ , in a random sample of size  $n$  from a continuous population with c.d.f.  $F(x)$  is:

$$F(m) = P(M \leq m) = n \int_{-\infty}^m [F(2m-x) - F(x)]^{n-1} \cdot f(x) dx$$

9. Let  $X_i$  ( $i = 1, 2, \dots, n$ ), be i.i.d. non-negative r.v.'s of continuous type. If  $M_n = X_{(n)} = \text{Max}(X_1, X_2, \dots, X_n)$ , and  $E(|X|) < \infty$ , then prove that

$$E(M_n) = E(M_{n-1}) + \int_0^\infty F^{n-1}(x) [1 - F(x)] dx$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

Hence find  $E(M_n)$  if  $X_i$ 's are i.i.d. exponential variates with parameter  $\lambda$ .

$$\text{Ans. } E(M_n) = \frac{1}{\lambda} \left[ 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} \right]$$

10. Show by means of an example that there may exist a r.v.  $X$  for which  $E(X)$  does not exist but  $E(X_{(r)})$  exists for some  $r$ ;

**Hint.** Let  $X_1, X_2, \dots, X_n$  be a random sample from the popln. with p.d.f.

$$f(x) = \frac{1}{x^2}, \quad 1 < x < \infty; \quad F(x) = 1 - \frac{1}{x}$$

$E(X)$  does not exist, but  $E(X_{(r)})$  exists for any  $r < n$ .

11. Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a population with p.d.f.

$$f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Show that  $Y_1 = X_{(1)}/X_{(2)}$ ,  $Y_2 = X_{(2)}/X_{(3)}$ , ..., ...,

$Y_{n-1} = X_{(n-1)}/X_{(n)}$  and  $Y_n = X_{(n)}$  are independently distributed and identify their distributions.

**Hint.** Proceed as in the hint to Q. No. 18 and 19.

12. Let  $X_1, X_2, X_3$  be a random sample of size 3 from exponential distribution with parameter  $\lambda$ . Show that  $Y_1 = X_{(3)} - X_{(2)}$  and  $Y_2 = X_{(2)}$  are independently distributed.

**Hint.**  $n = 3$ ; write joint p.d.f. of order statistics  $X_{(2)}$  and  $X_{(3)}$  and then transform to  $Y_1$  and  $Y_2$ .

13. Let  $X_1, X_2, \dots, X_{2m+1}$  be an odd-size random sample from a  $N(\mu, \sigma^2)$  population. Find the p.d.f. of the sample median and show that it is symmetric about  $\mu$ , and hence has the mean  $\mu$ .

14. A random sample of size  $n$  is drawn from an exponential population:

$$p(x) = \frac{1}{\theta} e^{-x/\theta}; \quad \theta > 0, \quad x \geq 0$$

(i) Obtain the p.d.f. of  $X_{(r)}$ .

(ii) Show that  $X_{(r)}$  and  $W_{rs} = X_{(s)} - X_{(r)}$ ,  $r < s$  are independently distributed.

(iii) Identify the distribution of  $W_1 = X_{(r+1)} - X_{(r)}$

[Gujrat Univ. M.Sc. (Stat.), 1991]

15. Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a uniform population with p.d.f.

$$f(x) = \begin{cases} 1, & \text{if } 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Show that :

(a)  $X_{(r)}$  is a  $\beta_1(r, n-r+1)$  variate.

(b)  $W_{rs} = X_{(s)} - X_{(r)}$  also has a Beta distribution which depends only on  $s-r$  and not on  $s$  and  $r$  individually.

$$\text{Ans. } f_{W_{rs}} = \frac{1}{\beta(s-r, n-s+r+1)} \cdot w_{rs}^{s-r-1} (1-w_{rs})^{n-s+r}; \quad 0 \leq w_{rs} \leq 1$$

16. In a random sample of size  $n$  from uniform  $U[0, 1]$  population, obtain the p.d.f. of  $W_{rs} = X_{(s)} - X_{(r)}$  and identify its distribution.

[Delhi Univ. M.Sc. (Stat.), 1987]

17. Obtain the p.d.f. of the range in a random sample of size 5 from the population with p.d.f.  $e^{-x}, x > 0$ . [Meerut Univ. M.Sc. (Stat.), 1993]

18. Let  $X_1, X_2, \dots, X_n$  be a random sample from continuous population with p.d.f.  $f(x) = \beta e^{-\beta x}; x \leq 0, \beta > 0$   
 $= 0, \text{ otherwise}$

(a) Show that  $X_{(r)}$  and  $X_{(s)} - X_{(r)}$  are independent for any  $s > r$ .

(b) Find the p.d.f. of  $X_{(r+1)} - X_{(r)}$

(c) Let  $Z_1 = nX_{(1)}, Z_2 = (n-1)(X_{(2)} - X_{(1)}),$

$$Z_3 = (n-2)(X_{(3)} - X_{(2)}), \dots, Z_n = (X_{(n)} - X_{(n-1)}) \quad \dots(*)$$

Show that  $(Z_1, Z_2, \dots, Z_n)$  and  $(X_1, X_2, \dots, X_n)$  are identically distributed.

**Hint.** (c) The joint p.d.f. of  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  is:

$$f(x_1, x_2, \dots, x_n) = n! f(x_1) f(x_2) \dots f(x_n) \\ = n! \beta^n \cdot e^{-\beta x_1} \cdot e^{-\beta x_2} \dots e^{-\beta x_n} \quad \dots(**)$$

Transformation (\*) gives:

$$X_{(1)} = \frac{Z_1}{n}; X_{(2)} = \frac{Z_1}{n} + \frac{Z_2}{n-1}; X_{(3)} = \frac{Z_1}{n} + \frac{Z_2}{n-1} + \frac{Z_3}{n-2}, \quad \dots(***)$$

$$X_{(n)} = \frac{Z_1}{n} + \frac{Z_2}{n-1} + \dots + \frac{Z_{n-1}}{2} + \frac{Z_n}{1}$$

$$J = \frac{\partial (X_{(1)}, X_{(2)}, \dots, X_{(n)})}{\partial (Z_1, Z_2, \dots, Z_n)} = \frac{1}{n!}$$

$$0 < X_{(1)} < X_{(2)} < \dots < X_{(n)} < \infty \Rightarrow 0 < Z_i < \infty; i = 1, 2, \dots, n.$$

Using (\*\*\*) and  $|J|$ , (\*\*) transforms to

$$g(z_1, z_2, \dots, z_n) = (\beta e^{-\beta z_1}) (\beta e^{-\beta z_2}) \dots (\beta e^{-\beta z_n}); 0 < z_i < \infty$$

$\Rightarrow Z_1, Z_2, \dots, Z_n$  are i.i.d. exponential variates with parameter  $\beta$ .

19. Let  $X_1, X_2, \dots, X_n$  be i.i.d. with p.d.f.

$$f(x) = \frac{1}{\sigma} \exp \left[ - \left( \frac{x-\theta}{\sigma} \right) \right]; x > 0 \\ = 0, \text{ otherwise}$$

Show that  $X_{(1)}, X_{(2)} - X_{(1)}, X_{(3)} - X_{(2)}, \dots, X_{(n)} - X_{(n-1)}$  are independent.

**Hint.**

$$Z_1 = X_{(1)}; Z_2 = X_{(2)} - X_{(1)}, \dots, Z_n = X_{(n)} - X_{(n-1)}$$

$$\Rightarrow X_{(1)} = Z_1, X_{(2)} = Z_1 + Z_2, \dots, X_{(n)} = Z_1 + Z_2 + \dots + Z_n$$

$$J = \frac{\partial (X_{(1)}, X_{(2)}, \dots, X_{(n)})}{\partial (Z_1, Z_2, \dots, Z_n)} = 1$$

As in above problem

$$g(z_1, z_2, \dots, z_n) = n! \prod_{i=1}^n f(x_i) |J|$$

$$= \left[ \frac{n}{\sigma} e^{-n(z_1 - \theta)/\sigma} \right] \cdot \left[ \frac{(n-1)}{\sigma} e^{-(n-1)z_2/\sigma} \right] \cdots \left[ \frac{2}{\sigma} e^{-2z_{n-1}/\sigma} \right] \times \left[ \frac{1}{\sigma} e^{-z/\sigma} \right]$$

$\Rightarrow Z_1, Z_2, \dots, Z_n$  are independently distributed.

20. Find the p.d.f. of  $i$ th order statistic.

Let  $X_1, X_2, \dots, X_n$  be i.i.d. with a distribution function

$$F(y) = \begin{cases} y^\alpha & \text{if } 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}; \quad \alpha > 0$$

Show that  $\frac{X_{(i)}}{X_{(n)}}$ ,  $i = 1, 2, \dots, n-1$  and  $X_{(n)}$  are independent.

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

21. Let  $x_{i1}, x_{i2}, \dots, x_{in}$ ,  $i = 1, 2, \dots, k$  be  $k$  random samples from  $R\left(\frac{1}{2}, 1\right)$  population. Find the distribution of  $U = x_{1(n)} \cdot x_{2(n)} \cdots x_{k(n)}$ , where  $x_{i(n)}$  is the maximum of  $i$ th sample.

(R = Rectangular population)

[Delhi Univ. M.Sc. (Stat.), 1989]

22. For the exponential distribution  $f(x) = e^{-x}$ ,  $x \geq 0$ , find the p.d.f. of the range  $W$  in a random sample of size  $n$  and show that

$$E(W) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}$$

Ans.  $g(w) = (n-1) e^{-w} (1-e^{-w})^{n-2}$ ;  $w \geq 0$ .

23. Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from  $U[a, b]$  population. Obtain the p.d.f.'s of (i)  $X_{(1)}$ , (ii)  $X_{(n)}$  and (iii) joint p.d.f. of  $X_{(1)}$  and  $X_{(n)}$ .

24. Let  $X_1, X_2$  be i.i.d. r.v.'s with p.d.f.

$$P(X_i = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots; \quad i = 1, 2$$

where  $\lambda > 0$ . Let  $M = \text{Max.}(X_1, X_2)$  and  $N = \text{Min.}(X_1, X_2)$

Find the marginal p.m.f.'s of  $M$  and  $N$ .

8-15. Truncated Distributions. Let  $X$  be a random variable with p.d.f. (or p.m.f.)  $f(x)$ . The distribution of  $X$  is said to be truncated at the point  $X = a$  if all the values of  $X \leq a$  are discarded. Hence the p.d.f. (or p.m.f.)  $g(\cdot)$  of the distribution, truncated at  $X = a$  is given by:

$$g(x) = \frac{f(x)}{P(X > a)}; \quad x > a \quad \dots(8.71)$$

$$= \frac{f(x)}{\sum_{x>a} f(x)}; \quad x > a \quad (\text{For discrete r.v.}) \quad \dots(8.71a)$$

$$= \frac{f(x)}{\int_a^{\infty} f(x) dx}; \quad x > a \text{ (For continuous r.v.)} \quad \dots(8.71b)$$

For the continuous r.v.  $X$ , the  $r$ th moment (about origin) for the truncated distribution is given by:

$$\mu_r' = E(X') = \int_a^{\infty} x' g(x) dx = \frac{\int_a^{\infty} x' f(x) dx}{\int_a^{\infty} f(x) dx} \quad \dots(8.72)$$

**Example 8.54** Let  $X \sim B(n, p)$ . Find the mean and variance of the binomial distribution truncated at  $X = 0$ .

**Solution.** Let  $f(x)$  be the p.m.f. of  $X \sim B(n, p)$  variate. Then the p.m.f.  $g(x)$  of the Binomial distribution truncated at  $X = 0$  is given by:

$$g(x) = \frac{f(x)}{P(X > 0)} = \frac{f(x)}{1 - P(X = 0)} = \frac{f(x)}{1 - f(0)} \\ = \frac{1}{1 - q^n} \cdot {}^n C_x p^x q^{n-x}; \quad x = 1, 2, \dots, n \quad \dots(*)$$

$$E(X) = \sum_{x=1}^n x g(x) = \frac{1}{1 - q^n} \sum_{x=1}^n x \cdot {}^n C_x p^x q^{n-x} \\ = \frac{1}{1 - q^n} \left[ \sum_{x=0}^n x \cdot {}^n C_x p^x q^{n-x} - 0 \right] \\ = np/(1 - q^n) \quad \dots(**)$$

$$E(X^2) = \frac{1}{1 - q^n} \sum_{x=1}^n x^2 \cdot {}^n C_x p^x q^{n-x} \\ = \frac{1}{1 - q^n} \left[ \sum_{x=0}^n x^2 \cdot {}^n C_x p^x q^{n-x} - 0 \right] \\ = \frac{1}{1 - q^n} [npq + n^2 p^2] \\ \left[ \because X \sim B(n, p); E(X^2) = \text{Var } X + (EX)^2 = npq + n^2 p^2 \right]$$

$$\therefore \text{Var}(X) = EX^2 - [E(X)]^2 \\ = \frac{1}{1 - q^n} \left[ npq + n^2 p^2 - \frac{n^2 p^2}{1 - q^n} \right]$$

**Example 8.55** Obtain the mean and variance of a standard Cauchy distribution truncated at both ends, with relevant range of variation as  $(-\beta, \beta)$ .

[Delhi Univ. M.Sc. (Stat.), 1987]

**Solution.** Let  $f(x)$  be the p.d.f. of standard Cauchy distribution. Then the

p.d.f.  $g(x)$  of the truncated distribution with relevant range of variation as  $(-\beta, \beta)$  is given by:

$$\begin{aligned} g(x) &= \frac{f(x)}{P(-\beta \leq X \leq \beta)} = \frac{f(x)}{\int_{-\beta}^{\beta} f(x) dx} \\ &= \frac{\frac{1}{\pi} \frac{1}{1+x^2}}{\frac{1}{\pi} \int_{-\beta}^{\beta} \frac{dx}{1+x^2}} = \frac{1}{1+x^2} \cdot \frac{1}{\left| \tan^{-1} x \right|_{-\beta}^{\beta}} \\ &= \frac{1}{2 \tan^{-1} \beta} \cdot \frac{1}{(1+x^2)} ; -\beta \leq x \leq \beta \quad \dots(*) \end{aligned}$$

$$\begin{aligned} \text{Mean} &= \int_{-\beta}^{\beta} x g(x) dx = \frac{1}{2 \tan^{-1} \beta} \int_{-\beta}^{\beta} \frac{x}{1+x^2} dx \\ &= 0 \quad [\because \text{Integrand is an odd function of } x] \end{aligned}$$

$$\begin{aligned} \text{Variance} &= \mu_2' - \mu_1'^2 = \mu_2' = \int_{-\beta}^{\beta} x^2 g(x) dx \\ &= \frac{2}{2 \tan^{-1} \beta} \int_0^{\beta} \frac{x^2}{1+x^2} dx = \frac{1}{\tan^{-1} \beta} \int_0^{\beta} \left( 1 - \frac{1}{1+x^2} \right) dx \\ &= \frac{1}{\tan^{-1} \beta} \left| x - \tan^{-1} x \right|_0^{\beta} = \frac{1}{\tan^{-1} \beta} (\beta - \tan^{-1} \beta) \\ &= \frac{\beta}{\tan^{-1} \beta} - 1 \end{aligned}$$

**Example 8-56** Consider a truncated standard normal distribution truncated at both ends with relevant range of variation as  $(A, B)$ . Obtain the p.d.f., mean, mode and variance of the truncated distribution.

[Delhi Univ. M.Sc.(Stat.), 1988]

**Solution.** Let  $Z \sim N(0, 1)$  with p.d.f.  $\varphi(z)$  and c.d.f.  $\Phi(z) = P(Z \leq z)$ . Then the p.d.f.  $g(\cdot)$  of the truncated normal distribution is given by:

$$g(z) = \frac{\varphi(z)}{P(A \leq Z \leq B)} = \frac{\varphi(z)}{\Phi(B) - \Phi(A)} = \frac{1}{k} \varphi(z) ; A \leq z \leq B \quad \dots(i)$$

where  $k = \Phi(B) - \Phi(A)$ .

$$\text{Mean} = \int_A^B z g(z) dz = \frac{1}{k} \int_A^B z \varphi(z) dz \quad \dots(ii)$$

We have  $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

$$\Rightarrow \frac{d}{dz} \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} (-z) = -z \varphi(z)$$

$$\Rightarrow \int z \varphi(z) dz = -\varphi(z) \quad \dots(*)$$

Substituting in (ii)

$$\text{Mean} = \frac{1}{k} \left| \int_A^B z \varphi(z) dz \right| = \frac{\varphi(A) - \varphi(B)}{\Phi(B) - \Phi(A)} = \mu' \text{ (say)} \quad \dots(iii)$$

$$\text{Mode} = \begin{cases} B & \text{if } A < 0, B < 0 \\ 0 & \text{if } A < 0, B > 0 \\ A & \text{if } A > 0, B > 0 \end{cases}$$

$$\begin{aligned} \text{Variance} &= \int_A^B z^2 g(z) dz - \mu'^2 \\ &= \frac{1}{k} \left[ \left| z(-\varphi(z)) \right|_A^B + \int_A^B \varphi(z) dz \right] - \mu'^2 \\ &= \frac{-1}{k} [ B \varphi(B) - A \varphi(A) - \{ \Phi(B) - \Phi(A) \} ] - \mu'^2 \\ &= 1 + \frac{A \varphi(A) - B \varphi(B)}{\Phi(B) - \Phi(A)} - \mu'^2 \end{aligned}$$

where  $\mu'$  is given in (iii).

### EXERCISE 8 (k)

1. Find the mean and variance of the truncated Poisson distribution with parameter  $\lambda$ , truncated at the origin.

**Ans. p.d.f.**

$$g(x) = \frac{1}{1-e^{-\lambda}} \left[ \frac{e^{-\lambda} \lambda^x}{x!} \right], x = 1, 2, 3, \dots; E(X) = \frac{\lambda}{1-e^{-\lambda}}; E(X^2) = \frac{\lambda + \lambda^2}{1-e^{-\lambda}}$$

2. Obtain the p.d.f. and the mean of truncated standard normal distribution, for positive values only.

$$\text{Ans. } g(z) = 2 \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} = 2\varphi(z); z > 0; E(Z) = \sqrt{2/\pi}$$

3. (a) Let  $X$  be normally distributed with mean  $\mu$  and variance  $\sigma^2$ . Truncate the density of  $X$  on the left at  $a$  and on the right at  $b$ , and then calculate the mean of the truncated distribution. [Note that if  $a = \mu - c$  and  $b = \mu + c$ , then the mean of the truncated distribution should equal  $\mu$ .]

[Delhi Univ. B.Sc. (Maths. Hons.), 1989]

(b) If  $X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ , find the mean of the conditional distribution of  $X$  given  $a \leq X \leq b$ .

**Hint.** In fact Problem in Part (a) is same as in Part (b), stated differently.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}; -\infty < x < \infty$$

Mean of truncated distribution is

$$\begin{aligned} \mu' &= \frac{\int_a^b xf(x) dx}{\int_a^b f(x) dx} = \frac{\int_a^b (x - \mu + \mu) f(x) dx}{\int_a^b f(x) dx} \\ &= \mu + \frac{\int_a^b (x - \mu) f(x) dx}{\int_a^b f(x) dx} = \mu + \sigma^2 \left[ \frac{f(a) - f(b)}{F(b) - F(a)} \right] \end{aligned}$$

where  $F(x) = P(X \leq x)$ , is the distribution function of  $X$ ,

$$\left\{ \begin{aligned} \because f'(x) &= f(x) \times \left[ -\left( \frac{x-\mu}{\sigma^2} \right) \right] \Rightarrow -\sigma^2 f'(x) = (x-\mu) f(x) \\ \Rightarrow \int_a^b (x-\mu) f(x) dx &= -\sigma^2 \left| f(x) \right|_{a}^{b} = \sigma^2 [f(a) - f(b)] \end{aligned} \right\}$$

4. A truncated Poisson distribution is given by the mass function

$$f(x) = \frac{1}{1-e^{-\lambda}} \cdot \frac{e^{-\lambda} \lambda^x}{x!}; x = 1, 2, 3, \dots, \lambda > 0$$

Find the m.g.f. and hence mean and variance of the distribution.

5. Consider the p.d.f.

$$g(x) = \frac{f(x)}{1 - F(x_0)}, \quad x > x_0 \quad \dots(*)$$

where  $f(x) = (\sqrt{2\pi} \cdot \sigma)^{-1} \cdot \exp \left[ -\frac{(x-\mu)^2}{2\sigma^2} \right], -\infty < \mu < \infty, \sigma > 0$

and  $F(x_0) = \int_{-\infty}^{x_0} f(u) du$

[(\*) is the p.d.f. of  $N(\mu, \sigma^2)$  distribution, truncated at the point  $x = x_0$ ].

Show that the first two raw moments can be expressed as:

$$\mu_1' = \mu + \lambda \sigma; \quad \mu_2' = \mu^2 + \lambda \sigma (x_0 + \mu) + \sigma^2$$

where  $\lambda [1 - F(x_0)] = f[(x_0 - \mu)/\sigma]$

6. Let  $X \sim \gamma(\alpha)$ . Obtain the p.d.f. of the truncated distribution, truncated at the point  $x_0$  and prove that  $\mu_r' = E X^r$  for the truncated gamma distribution  
 $= [E X^r \text{ for the untruncated } \gamma(\alpha) \text{ distribution}]$

$$\times \left[ \left\{ 1 - I_{x_0}(\alpha + r) \right\} / \left\{ 1 - I_{x_0}(\alpha) \right\} \right]$$

where  $I_{x_0}(\alpha) = \int_0^{x_0} \frac{e^{-x} x^{\alpha-1}}{\Gamma(\alpha)} dx$  (Incomplete Gamma Integral)

$$\text{Ans. } g(x) = \frac{1}{1 - I_{x_0}(\alpha)} \cdot \left( \frac{e^{-x} \cdot x^{\alpha-1}}{\Gamma(\alpha)} \right); \quad x > x_0$$

7. Explain the concept of 'Truncation'. For a standard normal distribution, truncated at both ends with relevant range of variation as  $[A, B]$ , obtain mean, variance and mean deviation about mean. [Delhi Univ. M.Sc. (Stat.), 1983]

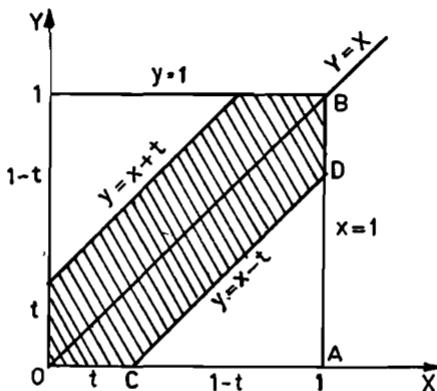
### ADDITIONAL EXERCISES ON CHAPTER VIII

1. If the random variable  $X$  has the density function  $f(x) = x/2, 0 < x < 2$ , find the  $r^{\text{th}}$  moment of  $X^2$ . Deduce that  $Z = X^2$  has the distribution  $g(y) = 1/4, 0 < y < 4$ .

2. Suppose that  $X$  is a random variable for which  $E(X) = \mu$  and  $V(X) = \sigma^2$ . Further suppose that  $Y$  is uniformly distributed over the interval  $(\alpha, \beta)$ . Determine  $\alpha$  and  $\beta$  so that  $E(X) = E(Y)$  and  $V(X) = V(Y)$ .

3. A boy and a girl agree to meet at a certain park between 4 and 5 P.M. They agree that the one arriving first will wait  $t$  hours,  $0 \leq t \leq 1$ , for the other to arrive. Assuming that the arrival times are independent and uniformly distributed, find the probability that they meet.

Hence obtain the probability that they meet if  $t = 10$  minutes.



**Hint.**  $p = P(|X - Y| \leq t) = \frac{\text{Shaded Area}}{\text{Total Area}}$

$$\begin{aligned} &= 2[\text{Area } OAB - \text{Area } CAD]/1 \times 1 \\ &= 1 - (1-t)^2 = 2t - t^2 \end{aligned}$$

**Ans.**  $t = 10$  minutes, Probability =  $11/36$

4. Let  $X \sim U[0, 1]$ . Find Corr.  $(X, Y)$ , where  $Y = X^n$

[Delhi Univ. B.A. (Hons.) (Spl. Course-Statistics), 1989]

5. (a) Let  $X_1, X_2, \dots, X_n$  be independent random variables having a common rectangular distribution over the interval  $[a, b]$ . Obtain the distribution of  $Y = \max. (X_1, X_2, \dots, X_n)$

(b) Let  $X \sim U\{0, 1, 2, \dots, r\}$ ,  $r = ab$ ,  $a > 1$ ,  $b < r$ , where  $a$  and  $b$  are positive integers. Show that the distribution of  $X$  coincides with  $U + V$ , where  $U$  and  $V$  are independent r.v.'s both with uniform distribution on appropriate subsets of  $\{0, 1, \dots, ab\}$ . (Indian Civil Services, 1988)

6. If  $X_1, X_2, \dots, X_n$  are mutually independent rectangular variates on  $[0, 1]$ , prove that the density function of  $X_1 \cdot X_2 \cdot \dots \cdot X_n$  is

$$f(x) = \begin{cases} \frac{(-\log x)^{n-1}}{(n-1)!}, & 0 < x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

7. (a) Let  $X_1$  and  $X_2$  be independent r.v.'s, each uniform on  $[0, 1]$ . Show that:

$Y_1 = \sqrt{-2 \log X_1} \cdot (\sin 2\pi X_2)$  and  $Y_2 = \sqrt{-2 \log X_1} \cdot \cos (2\pi X_2)$  are independent r.v.'s, and that each is  $N(0, 1)$ .

[This is known as Box and Muller transformation.]

(b) Given a sequence of independent r.v.'s  $X_1, X_2, \dots$  which are uniform on  $[0, 1]$ , produce a sequence of independent r.v.'s  $Y_1, Y_2, \dots$  that are  $N(0, 1)$  and independent.

(You may assume that the sum of two independent Normal distributions is itself Normally distributed.)

8. (a) Assume a random variable  $X$  has a standard normal distribution and let  $Y = X^2$

$$(i) \text{ Show that } F_Y(t) = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{t}} e^{-u^2/2} du, t \geq 0$$

(ii) Determine  $F_Y(t)$  when  $t < 0$  and describe the density function  $f_Y(t)$ .

(b) Let  $\Phi(x)$  be the standard normal distribution function and let

$$\Phi(x) = \int_{-\infty}^x \varphi(u) du$$

Show that

$$(i) \left( \frac{1}{x} - \frac{1}{x^2} \right) \varphi(x) \leq 1 - \Phi(x) \leq \frac{1}{x} \varphi(x), x > 0$$

$$(ii) \lim_{x \rightarrow \infty} \frac{x[1 - \Phi(x)]}{\varphi(x)} = 1$$

9. (a) If  $X_1$  and  $X_2$  are independent normal variates with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively, find the relation between  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$  so that

$$P(c_1 X_1 + c_2 X_2 < \alpha) = \gamma \text{ and } P(c_1 X_1 + c_2 X_2 < \beta) = \delta$$

(b) If  $X$  and  $Y$  are independent normal variates with equal means and standard deviations 9 and 12 respectively, and if

$$P[X + 2Y < 3] = P[2X - Y \geq 4],$$

determine the common mean of  $X$  and  $Y$ .

10. If  $X \sim N(0, 1)$  and  $A$  is constant, obtain the characteristic function of  $(X - A)^2$ . Hence or otherwise prove that

$$\kappa_r = 2^{r-1} (1 + rA^2) \cdot (r-1)!,$$

where  $\kappa_r$  is the  $r$ th cumulant of  $X$ .

$$\text{Ans. } \Phi_{(X-A)^2}(t) = (1 - 2it)^{-1/2} \cdot \exp\{it a^2/(1 - 2it)\}$$

11. (a) If  $X$  is a standard normal variate and  $\alpha$  is a small number, prove that  $P(X \leq \alpha) = \frac{1}{2} + \frac{8}{\sqrt{2\pi}} \left( \alpha - \frac{\alpha^3}{2^3 \cdot 3} + \frac{1}{2!} \cdot \frac{\alpha^5}{2^5 \cdot 5} - \dots \right)$

(b) Show that

$$\int_0^x e^{-t^2} dt = xe^{-x^2} \left[ 1 + \frac{1}{3} (2x^2) + \frac{1}{3 \cdot 5} (2x^2)^2 + \dots \right].$$

12. Let  $X$  be log-normal variate with p.d.f.

$$f(x, \mu, \sigma^2) = \frac{1}{x \sigma \sqrt{2\pi}} \exp \left\{ -(\log x - \mu)^2 / 2\sigma^2 \right\}, x > 0, \sigma > 0, |\mu| < \infty$$

Show that  $e^a X^b$  has a lognormal p.d.f.  $f(x, a + b \mu, b \sigma)$ . If  $X_1, X_2, \dots, X_n$  are  $n$  independent observations on  $X$ , then show that  $G = (X_1 X_2 \dots X_n)^{1/n}$  also has a log-normal p.d.f.  $f(x, \mu, \sigma/n)$ .

13. (a) The distribution

$$dF = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), -\infty < x < \infty.$$

is transformed by the transformation  $X = a \log_e(Y - b) + C$ . Find the distribution of  $Y$ . Evaluate the mean, the mode and the median of this distribution of  $(Y)$  and arrange them in order of magnitude when  $b > 0$ .

(b) If  $Y = a \log(X - b) + C$  has normal distribution with mean zero and unit variance, obtain the distribution of  $X$  and evaluate its mean, median and mode.

14. A standard variable  $X'$  is transformed to  $Y$  by the relation

$$X' = \frac{1}{c} [\log_{10} Y - a]$$

with  $m = e^{b^2 c^2}$  and  $\frac{1}{b} = \log_{10} e$ ; show that for the transformed variable

$$\beta_1 = m^2(m+3) - 4 \text{ and } \beta_2 = m^2(m^2 + 2m + 3) - 3$$

15. If  $X$  and  $Y$  are independent normal variates with zero expectations and variances  $\sigma_1^2$  and  $\sigma_2^2$ , show that  $Z = XY/\sqrt{X^2 + Y^2}$  is normal with variance  $\sigma_z^2 = [(1/\sigma_1^2) + (1/\sigma_2^2)]^{-1}$

16. If  $X_i; i = 1, 2, \dots, n$  is a random sample of size  $n$  from a normal population with mean  $\mu$  and variance  $\sigma^2$ , obtain the joint distribution of

$$U = \sum_{i=1}^n a_i X_i \text{ and } V = \sum_{i=1}^n b_i X_i$$

where  $a_i$ 's and  $b_i$ 's are arbitrary constants.

Hence or otherwise show that the necessary and sufficient condition that  $U$  and  $V$  are independent is that  $\sum a_i b_i = 0$ .

17. Let  $X_1, X_2$  and  $X_3$  be three independent normal variates with the same mean  $\mu$  and variance  $\sigma^2$ . Let

$$Y_1 = \frac{X_1 - X_2}{\sqrt{2}}, Y_2 = \frac{X_1 - 2X_2 + X_3}{\sqrt{6}}, \text{ and } Y_3 = \frac{X_1 + X_2 + X_3}{\sqrt{3}}$$

Show that  $Y_1, Y_2$  and  $Y_3$  are independent normal variates.

Show that

$$Y_1^2 + Y_2^2 = \sum_{i=1}^3 (X_i - \bar{X})^2, \text{ where } \bar{X} = \frac{X_1 + X_2 + X_3}{3}.$$

18. A random variable  $X$  has probability density function

$$f(x) = C \varphi(x), x \geq k$$

where  $k$  is a given number,  $C$  is a constant chosen to ensure that  $f(x)$  is a probability density function and

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2}x^2\right\}$$

If  $g(k) = \int_k^\infty \varphi(x) dx$ , show that the arithmetic mean and the variance of  $X$  are respectively  $\frac{\varphi(k)}{g(k)}$  and  $1 + \frac{\varphi(k)}{g(k)} \left\{ k - \frac{\varphi(k)}{g(k)} \right\}$

19. Three independent observations  $X_1, X_2, X_3$  are given from a univariate  $N(m, \sigma^2)$ . Derive the joint sampling distribution of :

- (a)  $U = X_1 - X_3$ ;
- (b)  $V = X_2 - X_3$
- (c)  $W = X_1 + X_2 + X_3 - 3m$

Deduce the p.d.f. of  $Z = U/V$ . Show that mode  $Z = 1/2$  and obtain the significance of this modal value. (Indian Civil Services, 1986)

20. Neyman's Contagious (Compound) Distribution. Let  $X \sim P(\lambda, y)$  where  $y$  itself is an observation of a variate  $Y \sim P(\lambda_1)$ . Find the unconditional distribution of  $X$  and show that its mean is less than its variance.

21. If  $X_1, X_2, \dots, X_n$  are independent random variables, having the probability law  $p(x_i) = \frac{e^{-\lambda_i} \lambda_i^{x_i}}{x_i!}$ ;  $i = 1, 2, \dots, n$ ,  $x_i = 0, 1, 2, \dots, \infty$

and if  $X = \sum_{i=1}^n X_i$  and  $\lambda = \sum_{i=1}^n \lambda_i$

then under certain conditions to be specified clearly,

$$P\left\{\frac{X-\lambda}{\sqrt{\lambda}} < \alpha\right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-t^2/2} dt \text{ as } n \rightarrow \infty.$$

22. Prove that

$$\frac{1}{n!} \int_{\lambda}^{\infty} e^{-x} x^n dx = \sum_{r=0}^n \frac{e^{-\lambda} \lambda^r}{r!}; n = 0, 1, 2, \dots$$

Using the above, write down the relation connecting the distribution functions of a Poisson and a Gamma variate.

23. A two-dimensional random variable  $(X, Y)$  has the joint p.d.f.,

$$f(x, y) = \frac{1}{\Gamma(\mu)\Gamma(\nu)} x^{\mu-1} (y-x)^{\nu-1} e^{-y}$$

in the  $x-y$  plane where  $0 < x < y < \infty$  and zero elsewhere. Show that the marginal distributions of  $X$  and  $Y$  are gamma distributions.

24. Starting from a suitable urn model, deduce the differential equation of the Pearsonian curves in the form

$$\frac{1}{y} \cdot \frac{dy}{dx} = \frac{a+x}{b_0 + b_1 x + b_2 x^2}$$

Also discuss the limitations of ranges in the solution of such differential equations.

25. Karl Pearson showed that the differential equation

$$\frac{d[f(x)]}{f(x)} = \frac{d-x}{a+bx+cx^2} dx$$

yields most of the important frequency curves when appropriate values of  $a, b, c$ , and  $d$  are chosen. Show that

(i) when  $d = 0$  and  $a = c = 0$  as well as  $b > 0$ , the differential equation yields exponential distribution,

(ii) when  $b = c = 0$  and  $a > 0$ , the differential equation yields normal distribution, and

(iii) when  $a = c = 0$ ,  $b > 0$  and  $d > -b$ , the differential equation yields gamma distribution.

26. Find the m.g.f. of the distribution with p.d.f.

$$f(x) = \left( \frac{\lambda}{2\pi x^3} \right)^{1/2} \exp \left[ \frac{-\lambda(x-\mu)^2}{2\mu^2 x} \right], x > 0, y > 0, \mu > 0 \quad \dots(*)$$

Also show that  $r$ th cumulant is given by:

$$K_r = 1.3.5 \dots (2r-3) \mu^{2r-3} \lambda^{1-r}$$

(\*) is the p.d.f. of *Standard Inverse Gaussian distribution*.

27. Obtain  $M_X(t)$ , when  $X \sim P(\lambda)$ . Find the limit as  $\lambda \rightarrow \infty$  of m.g.f. of  $(X - \lambda)/\sqrt{\lambda}$  and interpret the result in the context of C.L.T. Also prove that:

$$\lim_{\lambda \rightarrow \infty} \sum_{k=a}^{\beta} \left[ \frac{e^{-\lambda} \cdot \lambda^k}{k!} \right] = \frac{1}{\sqrt{2\pi}} \int_a^{\beta} \exp \left( -\frac{1}{2} u^2 \right) du, \\ \alpha = \lambda + a\sqrt{\lambda}, \quad \beta = \lambda + b\sqrt{\lambda}$$

Show that  $2X$  is not a Poisson variate. Give a set of conditions under which  $X + Y$  too is a Poisson variate. (Indian Civil Services, 1983)

28. The random variables  $X_k$  ( $k = 1, 2, \dots$ ) are independent and have the Cauchy distribution with p.d.f.

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, -\infty < x < \infty. \text{ Let } Y_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Examine whether the sequence  $\{Y_n\}$  obeys the weak law of large numbers.

29. Let  $X_1, X_2, \dots, X_n, \dots$  be independent Bernoulli variates such that

$$P(X_k = 1) = p_k, \quad P(X_k = 0) = q_k = (1 - p_k), \quad k = 1, 2, \dots, n, \dots$$

Examine whether the sequence  $\left\{ \sum_{i=1}^n X_i \right\}$  follows the central limit theorem.

### OBJECTIVE TYPE QUESTIONS

1. Choose the correct answer from B and match it with each item in A.

A

B

- |   |                                  |
|---|----------------------------------|
| (a) $\beta_2$ for a Normal distribution               | (1) $3\sigma^4$                  |
| (b) $\beta_1$ for a Normal distribution               | (2) 0                            |
| (c) $\mu_3$ for a Normal distribution                 | (3) 3                            |
| (d) $\mu_4$ for a Normal distribution                 | (4) $e^{i\mu t - t^2\sigma^2/2}$ |
| (e) Characteristic function for Normal distribution   | (5) $\frac{4}{3}\sigma$          |
| (f) Moment generating function of Normal distribution | (6) 0                            |
| (g) Mode of Normal distribution                       | (7) $e^{\mu t + t^2\sigma^2/2}$  |
| (h) Mean deviation from mean for Normal distribution  | (8) $\mu$                        |

11. Match the distributions:

- |                              |   |
|------------------------------|---|
| (a) Uniform distribution     | (1) $f(x) = \frac{1}{\lambda[(x-\mu)^2 + \lambda^2]}, -\infty < x < \infty$ |
| (b) Normal distribution      | (2) $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\infty < x < \infty$         |
| (c) Exponential distribution | (3) $f(x) = \frac{1}{b-a}, a \leq x \leq b$                                 |
| (d) Beta distribution        | (4) $f(x) = \frac{1}{B(m, n)} x^{m-1} (1-x)^{n-1}, 0 \leq x \leq 1$         |
| (e) Cauchy distribution      | (5) $f(x) = \frac{1}{a} e^{- x/a }, x \geq 0$                               |

III. If  $X_i$  ( $i = 1, 2, 3, \dots, n$ ) are independent  $N(0, 1)$ , write (without proof), the distribution of

- (i)  $X_1 - 2X_2 + X_3$ ,
- (ii)  $\frac{X_2}{X_3}$ ,
- (iii)  $\sum_{i=1}^n X_i$
- (iv)  $\frac{X_1^2}{X_1^2 + X_2^2}$ ,
- (v)  $\frac{X_1^2}{X_2^2 + X_3^2}$ ,
- (vi)  $\frac{X_i}{X_j}, (i \neq j)$
- (vii)  $\sum_{i=1}^n X_i^2$
- (viii)  $\frac{X_i^2}{X_j^2}, i \neq j$

IV. In each case, specify the distribution for which:

- (i) Moments do not exist.
- (ii) Mean = variance.

(iii) Mean < variance

(iv) Mean > variance.

(v)  $\phi(t) = e^{it - t^2}$

(vi)  $\phi(t) = e^{-|t|}$

V. State the conditions under which

(i) Binomial distribution,

(ii) Poisson distribution

tends to Normal distribution.

VI. (i) Give two examples of variates which you expect to be distributed normally.

(ii) Give two examples of variates which you expect to be distributed exponentially.

VII. State which of the following statements are TRUE and which are FALSE. In case of false statements, give the correct statement.

(i) For normal distribution mean deviation about mean is greater than quartile deviation.

(ii)  $X$  is a random variable following Cauchy distribution, for which mean does not exist but variance exists.

(iii) In case of normal distribution  $\beta_1 = 3$ ,  $\beta_2 = 0$ ,

(iv) If  $X$  and  $Y$  are two independent normal variates, then  $X - Y$  is also a normal variate.

(v) Binomial distribution tends to normal distribution as  $n \rightarrow \infty$ .

(vi) For normal distribution, mean = mode = median.

(vii) It is possible to reduce every normal distribution to the standard normal distribution by a transformation.

(viii) In uniform distribution, the percentile points are equi-spaced.

(ix) Normal distribution is symmetrical only for some specified values of the mean and variance.

(x) Normal distribution can be obtained as a limiting case of Poisson distribution with the parameter  $\lambda \rightarrow \infty$ .

VIII. Give the correct answer to each of the following :

(i) The mean and variance of Normal distribution

(a) are same, (b) cannot be same, (c) are sometimes equal, (d) are equal in the limiting case, as  $n \rightarrow \infty$ .

(ii) The mean and variance of Gamma distribution

(a) are same, (b) cannot be same, (c) are sometimes equal, (d) are equal in the limiting case, as  $n \rightarrow \infty$ .

(iii)  $X$  is normally distributed with zero mean and unit variance. The variance of  $X^2$  is

(a) 0, (b) 1, (c) 2, (d) 4

(iv) The points of inflexion of Normal curve are

(a)  $m \pm \sigma$ , (b)  $m \pm 2\sigma$ , (c)  $m \pm 3\sigma$ , (d)  $m \pm \frac{2}{3}\sigma$

- (v) The moment generating function of gamma distribution is  
 (a)  $(1+t)^\lambda$ , (b)  $(1-t)^\lambda$ , (c)  $(1-t)^{-\lambda}$ , (d)  $(1+t)^{-\lambda}$
- (vi) The characteristic function of Cauchy distribution is  
 (a)  $e^{-t}$ , (b)  $e^{-|t|}$ , (c)  $e^t$ , (d)  $e^{|t|}$
- (vii) Area to the right of the point  $x_1$  is 0.6 and to the left of the point  $x_2$ , is 0.7. Which is the correct:  
 (i)  $x_1 > x_2$ , (ii)  $x_1 < x_2$  or (iii)  $x_1 = x_2$ ?
- (viii) The standard normal distribution is represented by  
 (a)  $N(0, 0)$ , (b)  $N(1, 1)$ , (c)  $N(0, 1)$ , (d)  $N(1, 0)$ .
- (ix) For a normal distribution, quartile deviation, mean deviation, standard deviation are in the ratio  
 (a)  $\frac{4}{5} : \frac{2}{3} : 1$ , (b)  $\frac{2}{3} : \frac{4}{5} : 1$ , (c)  $1 : \frac{4}{5} : \frac{2}{3}$ , (d)  $\frac{2}{3} : 1 : \frac{4}{5}$
- (x) The normal distribution is a limiting form of binomial distribution if  
 (a)  $n \rightarrow \infty$ ,  $p \rightarrow 0$ , (b)  $n \rightarrow 0$ ,  $p \rightarrow q$ , (c)  $n \rightarrow \infty$ ,  $p \rightarrow n$ , (d)  $n \rightarrow \infty$  and neither  $p$  nor  $q$  is small.
- (xi) Normal curve is  
 (a) very flat, (b) bell shaped symmetrical about mean, (c) very peaked, (d) smooth.
- (xii) The normal distribution is a limiting case of Poisson's when  
 (a)  $\lambda \rightarrow 0$ , (b)  $\lambda \rightarrow \sigma$ , (c)  $\lambda \rightarrow \infty$ , (d)  $\lambda < \sigma$ .
- (xiii) In a normal curve the number of observations less than mean are included in the range  
 (a)  $\bar{x} \pm 3\sigma$ , (b)  $\bar{x} \pm 1.96$ , (c)  $\bar{x} \pm 2\sigma$ , (d)  $\bar{x} \pm 0.67\sigma$
- (xiv) If  $X$  is a standard normal variate, then  $\frac{1}{2}X^2$  is a  
 (a) Gamma variate with parameter  $\frac{1}{2}$ , (b) a normal variate, (c) a Poisson variate.
- (xv) The range of the beta variate is  
 (a)  $(0, \infty)$ , (b)  $(-\infty, \infty)$ , (c)  $(0, 1)$ , (d)  $(-1, +1)$

**IX. Fill in the blanks :—**

- The mean deviation of normal distribution is...
- The p.d.f. of Gamma distribution is...
- The relationship between Beta distributions of first and second kind is...
- The normal distribution is a limiting form of binomial distribution if...
- Mean = variance for ... distribution (continuous).
- For the normal distribution :

$$\beta_1 = \dots$$

$$\beta_2 = \dots$$

$$\text{Mean deviation} = \dots$$

Quartile deviation = ...

(vii) The characteristic function of a Gamma distribution is ...

(viii) The points of inflexion for a normal curve are ...

(ix) For the normal distribution with variance  $\sigma^2$ ,

$$\mu_{2r} = \dots$$

$$\mu_{2r+1} = \dots$$

(x) A normal distribution is completely specified by the parameters...

(xi) For normal distribution

$$\text{S.D. : M.D. : Q.D. :: ... : ... : ...}$$

(xii) If  $X \sim N(\mu, \sigma^2)$  then

$$P(\mu - \sigma < X < \mu + \sigma) = \dots$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \dots$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = \dots$$

(xiii) If  $X$  is a random variable with distribution function  $F$  then  $F(X)$  has ... distribution.

(xiv) If  $X \sim N(0, 1)$ , then  $X^2/2$  has ... distribution with parameter ...

X. Random variables  $X_i$  are independent and all of them have the same distribution defined by

$$f(x) = \frac{1}{\sqrt{8\pi}} \exp \left\{ -(x-1)^2/8 \right\}, \quad -\infty < x < \infty$$

Find the distribution of

$$(i) \sum_{i=1}^{10} X_i / 10 \text{ and } (ii) X_1 - 2X_2 + X_3$$

$$\text{Ans. (i)} \quad N\left(1, \frac{2}{5}\right), \quad \text{(ii)} \quad N(0, 24)$$

XI. The random variables  $X_i, i = 1, 2, \dots$  are independent and all of them have the same distribution defined by

$$f(x) = \frac{1}{\sqrt{18\pi}} e^{-(x-1)^2/18}; \quad -\infty < x < \infty.$$

Find the distribution of

$$(a) \quad \frac{1}{5} \sum_{i=1}^5 X_i.$$

$$(b) \quad 3X_1 - X_2 + 2X_3.$$

XII. Find the mean and standard deviation of a probability distribution whose frequency function is given by

$$f(x) = Ce^{-(1/24)(x^2 - 6x + 9)}, \quad -\infty < x < \infty$$

where  $C$  is a constant.

[Delhi Univ. B.A. (Stat. Hons.), 1986]

$$\text{Ans. Mean } 3, \sigma^2 = 12, C = \frac{1}{\sqrt{24\pi}}$$

XIII. (a) If  $f(x) = ke^{-3x^2 + 6x}$ ,  $-\infty < x < \infty$ ,  
obtain the values of  $k$ ,  $\mu$  and  $\sigma^2$ .

Ans.  $k = \sqrt{(3/\pi)} e^{-3}$ ;  $\mu = 1$ ,  $\sigma^2 = \frac{1}{6}$

(b) If  $X$  is a normal variate with mean  $\mu$  and variance  $\sigma^2$ , find the distribution of  $Y = aX + b$ .

Ans.  $Y \sim N(a\mu + b, a^2\sigma^2)$ .

(c) If  $X$  is distributed normally with mean  $\mu$  and standard deviation  $\sigma$ , write down the distribution of  $U = 2X - 3$  and find the mean and variance of  $U$ .

Ans.  $U \sim N(2\mu - 3, 4\sigma^2)$

XIV. If  $X_1$  is normally distributed with a mean 10 and variance 16 and  $X_2$  is normally distributed with a mean 10 and variance 15 and if  $W = X_1 + X_2$ , what will be the values of the two parameters of the distribution of the variate  $W$ ? (Assume that  $X_1$  and  $X_2$  are independent).

(c) Let  $X$  and  $Y$  be independent and normally distributed as  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$ . Find the mean and variance of  $\frac{1}{2}(X + Y)$

XV. Write a note on the role of Normal distribution in Statistics.

XVI. If  $X_i$ , ( $i = 1, 2, 3, \dots, n$ ) are i.i.d. standard Cauchy variates, write the distribution of  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .

Ans. Standard Cauchy.

XVII. Is the sum of two independent Cauchy variates a Cauchy variate?

If  $X_i$ , ( $i = 1, 2, 3, 4$ ) are independent standard normal variates, what is the distribution of  $\frac{X_1}{X_2} + \frac{X_3}{X_4}$ ?

Ans. Cauchy.

XVIII. "The role of Cauchy distribution in statistical theory often lies in providing counter examples." Elucidate.

XIX. Write a note on the role of Central Limit Theorem in Statistics.

XX. If  $X_i$ , ( $i = 1, 2, 3, 4$ ) are i.i.d.  $N(0, 1)$ , write the distribution of:

(i)  $X_1 - X_2$  (ii)  $X_1 + X_2$  (iii)  $\frac{X_1}{X_2}$  (iv)  $\frac{X_1}{X_2} + \frac{X_3}{X_4}$

(v)  $X_1 - X_2 + X_3 - X_4$  (vi)  $\frac{1}{2}X_1^2$  (vii)  $\frac{1}{2}(X_2^2 + X_3^2)$

(viii)  $\frac{X_1^2}{X_1^2 + X_2^2}$  (ix)  $\frac{X_3^2}{X_4^2}$  (x)  $\frac{X_1^2}{X_2^2 + X_3^2}$ .

Ans. (i)  $N(0, 2)$ ; (ii)  $N(0, 2)$ ; (iii) Standard Cauchy; (iv) Cauchy; (v)  $N(0, 4)$ ; (vi)  $\gamma(\frac{1}{2})$ ; (vii)  $\gamma(1)$ ; (viii)  $\beta_1(\frac{1}{2}, \frac{1}{2})$ ; (ix)  $\beta_2(\frac{1}{2}, \frac{1}{2})$ ; (x)  $\beta_2(\frac{1}{2}, 1)$ .

## CHAPTER NINE

# *Curve Fitting and Principle of Least Squares*

**9.1. Curve Fitting.** Let  $(x_i, y_i); i = 1, 2, \dots, n$  be a given set of  $n$  pairs of values,  $X$  being independent variable and  $Y$  the dependent variable. The general problem in curve fitting is to find, if possible, an analytic expression of the form  $y = f(x)$ , for the functional relationship suggested by the given data.

Fitting of curves to a set of numerical data is of considerable importance—theoretical as well as practical. Theoretically it is useful in the study of correlation and regression, e.g., lines of regression can be regarded as fitting of linear curves to the given bivariate distribution (c.f. § 10.8.1). In practical statistics it enables us to represent the relationship between two variables by simple algebraic expressions, e.g., polynomials, exponential or logarithmic functions. Moreover, it may be used to estimate the values of one variable which would correspond to the specified values of the other variable.

**9.1.1. Fitting of a straight line.** Let us consider the fitting of a straight line

$$Y = a + bX \quad \dots(9.1)$$

to a set of  $n$  points  $(x_i, y_i); i = 1, 2, \dots, n$ . Equation (9.1) represents a family of straight lines for different values of the arbitrary constants 'a' and 'b'. The problem is to determine 'a' and 'b' so that the line (9.1) is the line of "best fit".

The term 'best fit' is interpreted in accordance with Legendre's principle of least squares which consists in minimising the sum of the squares of the deviations of the actual values of  $y$  from their estimated values as given by the line of best fit.

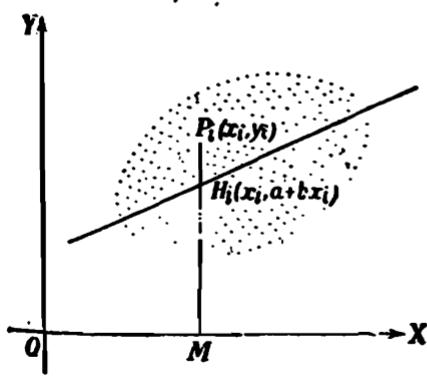
Let  $P_i(x_i, y_i)$  be any general point in the scatter diagram (§ 10.2). Draw  $P_i M \perp$  to  $x$ -axis meeting the line, (9.1) in  $H_i$ . Abscissa of  $H_i$  is  $x_i$  and since  $H_i$  lies on (9.1), its ordinate is  $a + bx_i$ . Hence the co-ordinates of  $H_i$  are  $(x_i, a + bx_i)$ .

$$P_i H_i = P_i M - H_i M$$

$$= y_i - (a + bx_i),$$

is called the *error of estimate* or the *residual for  $y_i$* .

According to the principle of



least squares, we have to determine  $a$  and  $b$  so that

$$E = \sum_{i=1}^n P_i H_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

is minimum. From the principle of maxima and minima, the partial derivatives of  $E$ , with respect to (w.r.t.)  $a$  and  $b$  should vanish separately, i.e.,

$$\frac{\partial E}{\partial a} = 0 = -2 \sum_{i=1}^n (y_i - a - bx_i) \quad \text{and} \quad \frac{\partial E}{\partial b} = 0 = -2 \sum_{i=1}^n x_i (y_i - a - bx_i) \quad \dots(9-2)$$

$$\Rightarrow \sum_{i=1}^n y_i = na + b \sum_{i=1}^n x_i \quad \text{and} \quad \sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 \quad \dots(9-2a)$$

Equations (9-2) and (9-2a) are known as the *normal equations* for estimating  $a$  and  $b$ .

All the quantities  $\sum_{i=1}^n x_i$ ,  $\sum_{i=1}^n x_i^2$ ,  $\sum_{i=1}^n y_i$  and  $\sum_{i=1}^n x_i y_i$ , can be obtained from the given set of points  $(x_i, y_i)$ ;  $i = 1, 2, \dots, n$  and the equations (9-2a) can be solved for  $a$  and  $b$ . With the values of  $a$  and  $b$  so obtained, equation (9-1) is the line of best fit to the given set of points  $(x_i, y_i)$ ;  $i = 1, 2, \dots, n$ .

**Remark.** The equation of the line of best fit of  $y$  on  $x$  is obtained on eliminating  $a$  and  $b$  in (9-1) and (9-2a) and can be expressed in the determinant form as follows:

$$\begin{vmatrix} Y & X & 1 \\ \sum y_i & \sum x_i & n \\ \sum x_i y_i & \sum x_i^2 & \sum x_i \end{vmatrix} = 0 \quad \dots(9-2b)$$

### 9-1-2. Fitting of second degree parabola. Let

$$Y = a + bX + cX^2 \quad \dots(9-3)$$

be the second degree parabola of best fit to set of  $n$  points  $(x_i, y_i)$ ;  $i = 1, 2, \dots, n$ . Using the principle of least squares, we have to determine the constants  $a$ ,  $b$  and  $c$  so that

$$E = \sum_{i=1}^n (y_i - a - bx_i - cx_i^2)^2$$

is minimum.

Equating to zero the partial derivatives of  $E$  with respect to  $a$ ,  $b$  and  $c$  separately, we get the normal equations for estimating  $a$ ,  $b$  and  $c$  as

$$\left. \begin{aligned} \frac{\partial E}{\partial a} &= 0 = -2 \sum (y_i - a - bx_i - cx_i^2) \\ \frac{\partial E}{\partial b} &= 0 = -2 \sum x_i (y_i - a - bx_i - cx_i^2) \\ \frac{\partial E}{\partial c} &= 0 = -2 \sum x_i^2 (y_i - a - bx_i - cx_i^2) \end{aligned} \right\} \quad \dots(9-4)$$

$$\Rightarrow \begin{aligned} \sum y_i &= na + b \sum x_i + c \sum x_i^2 \\ \sum x_i y_i &= a \sum x_i + b \sum x_i^2 + c \sum x_i^3 \\ \sum x_i^2 y_i &= a \sum x_i^2 + b \sum x_i^3 + c \sum x_i^4, \end{aligned} \quad \dots(9.4a)$$

summation taken over  $i$  from 1 to  $n$ .

For given set of points  $(x_i, y_i); i = 1, 2, \dots, n$ , equations (9.4a) can be solved for  $a$ ,  $b$  and  $c$ , and with these values of  $a$ ,  $b$  and  $c$ , (9.3) is the parabola of best fit.

**Remark.** Eliminating  $a$ ,  $b$  and  $c$  in (9.3) and (9.4a), the parabola of best fit of  $y$  on  $x$  is given by

$$\left| \begin{array}{cccc} Y & X^2 & X & 1 \\ \sum y_i & \sum x_i^2 & \sum x_i & n \\ \sum x_i y_i & \sum x_i^3 & \sum x_i^2 & \sum x_i \\ \sum x_i^2 y_i & \sum x_i^4 & \sum x_i^3 & \sum x_i^2 \end{array} \right| = 0 \quad \dots(9.4b)$$

### 9.1.3. Fitting of Polynomial of $k$ th Degree. If

$$Y = a_0 + a_1 X + a_2 X^2 + \dots + a_k X^k \quad \dots(9.5)$$

is the  $k^{th}$  degree polynomial of best fit to the set of points  $(x_i, y_i); i = 1, 2, \dots, n$ , the constants  $a_0, a_1, a_2, \dots, a_k$  are to be obtained so that

$$E = \sum_{i=1}^n (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_k x_i^k)^2$$

is minimum. Thus the normal equations for estimating  $a_0, a_1, \dots, a_k$  are obtained on equating to zero the partial derivatives of  $E$  w.r.t.  $a_0, a_1, \dots, a_k$  separately, i.e.,

$$\left. \begin{aligned} \frac{\partial E}{\partial a_0} &= 0 = -2 \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_k x_i^k) \\ \frac{\partial E}{\partial a_1} &= 0 = -2 \sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_k x_i^k) \\ \frac{\partial E}{\partial a_k} &= 0 = -2 \sum x_i^k (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_k x_i^k) \end{aligned} \right\} \quad \dots(9.6)$$

$$\Rightarrow \left. \begin{aligned} \sum y_i &= n a_0 + a_1 \sum x_i + a_2 \sum x_i^2 + \dots + a_k \sum x_i^k \\ \sum x_i y_i &= a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 + \dots + a_k \sum x_i^{k+1} \\ \sum x_i^k y_i &= a_0 \sum x_i^k + a_1 \sum x_i^{k+1} + a_2 \sum x_i^{k+2} + \dots + a_k \sum x_i^{2k}, \end{aligned} \right\} \quad \dots(9.6a)$$

summation extended over  $i$  from 1 to  $n$ . These are  $(k+1)$  equations in  $(k+1)$  unknowns  $a_0, a_1, a_2, \dots, a_k$  and can be solved with the help of algebra.

**Remark.** It has been found that in all the above cases, the values of the second order derivatives, viz.,  $\frac{\partial^2 E}{\partial a_0^2}, \frac{\partial^2 E}{\partial a_1^2}, \dots$  come out to be positive at the

points  $a_0, a_1, \dots, a_k$ , the solutions of the 'normal equations'. Hence they provide minima of  $E$ . For proof see Remark 1 to § 10.7.1—Lines of Regression.

**Example 9.1.** Fit a straight line to the following data.

$X :$	1	2	3	4	6	8
$Y :$	2.4	3	3.6	4	5	6

**Solution.** Let the line be  $Y = a + bX$

$X$	$Y$	$X^2$	$\dots$	$XY$
1	2.4	1		2.4
2	3.0	4		6.0
3	3.6	9		10.8
4	4.0	16		16.0
6	5.0	36		30.0
8	6.0	64		48.0
Total	24	130		113.2

Using normal equations (9.2a), we get

$$24 = 6a + 24b \text{ and } 113.2 = 24a + 130b$$

Solving these equations, we get  $a = 1.976$  and  $b = 0.506$ .

**Example 9.2.** Fit a parabola of second degree to the following data :

$X :$	0	1	2	3	4
$Y :$	1	1.8	1.3	2.5	6.3

(Delhi Univ. B.Sc., Oct. 1992)

**Solution.** Let  $Y = a + bX + cX^2$  be the second degree parabola.

	$X$	$Y$	$X^2$	$X^3$	$X^4$	$XY$	$X^2Y$
	0	1.0	0	0	0	0	0
	1	1.8	1	1	1	1.8	1.8
	2	1.3	4	8	16	2.6	5.2
	3	2.5	9	27	81	7.5	22.5
	4	6.3	16	64	256	25.2	100.8
Total	10	12.9	30	100	354	37.1	130.3

Using normal equations (9.4), we get

$$12.9 = 5a + 10b + 30c ; 37.1 = 10a + 30b + 100c ;$$

$$130.3 = 30a + 100b + 354c$$

Solving these equations, we get  $a = 1.42$ ,  $b = -1.07$  and  $c = 0.55$ . Thus the required equation of the second degree parabola is

$$Y = 1.42 - 1.07X + 0.55X^2$$

**Remark.** If the values which  $X$  and  $Y$  take are large, the calculation of  $\Sigma x$ ,  $\Sigma x^2$ ,  $\Sigma xy$ , ..., becomes quite tedious and the solution of the normal equations, is also quite cumbersome. In this case arithmetic is reduced to a great

extent by suitable change of origin in  $X$  or (and) in  $Y$ .

**9.1.4. Change of origin.** Let us suppose that the values of  $X$  are given to be equidistant at an interval of  $h$ , i.e.,  $X$  takes the values, (say),  $a, a+h, a+2h, \dots$ . If  $n$  is odd, i.e.,  $n = 2m+1$  (say), we take

$$U = \frac{X - (\text{middle term})}{\text{Interval}} = \frac{X - (a + mh)}{h}$$

Now  $U$  takes the values  $-m, -(m-1), \dots, -1, 0, 1, \dots, (m-1), m$ , so that  $\sum U = \sum U^3 = 0$ .

If  $n$  is even, i.e.,  $n = 2m$  (say), then there are two middle terms, viz.,  $m$ th and  $(m+1)$ th terms which are  $a+(m-1)h$  and  $a+mh$ . In this case, we take

$$\begin{aligned} U &= \frac{X - (\text{mean of two middle terms})}{\frac{1}{2}(\text{interval})} = \frac{X - [a + \frac{1}{2}(2m-1)h]}{\frac{1}{2}(h)} \\ &= \frac{2X - 2a - (2m-1)h}{h} \end{aligned} \quad \dots(9.7)$$

Now for  $X = a, a+h, \dots, a+(2m-1)h$ ;  $U$  takes the values  $-(2m-1), -(2m-3), \dots, -3, -1, 1, 3, \dots, (2m-3), (2m-1)$ .

Again we see that  $\sum U = \sum U^3 = 0$ .

**Example 9.3.** The weights of a calf taken at weekly intervals are given below. Fit a straight line using the method of least squares and calculate the average rate of growth per week.

Age ( $X$ ) : 1 2 3 4 5 6 7 8 9 10  
Weight ( $Y$ ) : 52.5 58.7 65.0 70.2 75.4 81.1 87.2 95.5 102.2 108.4

**Solution.** Let the variables age and weight be denoted by  $X$  and  $Y$  respectively.

Here  $n = 10$ , i.e., even and the values of  $X$  are equidistant at an interval of unity, i.e.,  $h = 1$ . Thus we take

$$U = \frac{X - ((5+6)/2)}{\frac{1}{2}} = 2X - 11.$$

Let the least-square line of  $Y$  on  $U$  be  $Y = a + bU$ .

The normal equations for estimating  $a$  and  $b$  are

$$\Sigma Y = na + b\Sigma U \quad \text{and} \quad \Sigma UY = a\Sigma U + b\Sigma U^2$$

$X$	$Y$	$U$	$U^2$	$UY$
1	52.5	-9	81	-472.5
2	58.7	-7	49	-410.9
3	65.0	-5	25	-325.0
4	70.2	-3	9	-210.6
5	75.4	-1	1	-75.4
6	81.1	1	1	81.1
7	87.2	3	9	261.6

8	95.5	5	25	477.5
9	102.2	7	49	715.4
10	108.4	9	81	975.6
Total	796.2	0	330	1016.8

Thus the normal equations are

$$796.2 = 10a + 0 \times b \quad \text{and} \quad 1016.8 = a \times 0 + 330b,$$

which give  $a = 79.62$  and  $b = \frac{1016.8}{330} = 3.08$  (approx).

∴ The least square line of  $Y$  on  $U$  is

$$Y = 79.62 + 3.08U$$

Hence the line of best fit of  $Y$  on  $X$  is

$$Y = 79.62 + 3.08(2X - 11) \Rightarrow Y = 45.74 + 6.16X$$

The weights of the calf (as given by the line of best fit  $Y = A + BX$ ) after 1, 2, 3, ... weeks are  $A + B$ ,  $A + 2B$ ,  $A + 3B$ , ..., respectively. Hence the average rate of growth per week is  $B$  units, i.e., 6.16 units.

### EXERCISE 9 (a)

1. (a)  $(x_i, y_i); i = 1, 2, \dots, n$ , give the co-ordinates of  $n$  points in a plane. It is proposed to fit a straight line  $Y = aX + b$  to those points such that the sum of the squares of the perpendiculars from those  $n$  points to the line is a minimum. Find the constants  $a$  and  $b$ . Use the above method to fit a straight line to the following points :

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

Ans.  $Y = 0.72 + 1.33X$

- (b) Fit a straight line of the form  $Y = AX + B$  to the following data:

$X:$	0	5	10	15	20	25	30
$Y:$	10	14	19	25	31	36	39

2. Show that the line of best fit to the following data is given by

$$Y = -0.5X + 8$$

$X:$	6	7	7	8	8	8	9	9	10
$Y:$	5	5	4	5	4	3	4	3	3

3. (a) How do you define the term "line of best fit". Give the normal equations generally used to obtain such a line. Fit a straight line and parabolic curve to the following data :  $X : 1.01, 5.2, 0.02, 5.3, 0.03, 5.4, 0$

$Y:$	1.1	1.3	1.6	2.6	2.7	3.4	4.1
------	-----	-----	-----	-----	-----	-----	-----

Ans.  $Y = 1.04 - 0.20X + 0.24X^2$

- (b) Fit a straight line to the following data. Plot the observed and the ex-

(b) Fit a straight line to the following data. Plot the observed and the expected values in a graph and examine whether the straight line gives an adequate fit.

$x \dots$	1	2	3	4	5	6	7	8
$y \dots$	55	46	40	38	33	30	29	30

4. An experiment is conducted to verify the law of falling under gravity expressed by  $S = \frac{1}{2}gt^2$

where  $S$  is the distance fallen at time  $t$  and  $g$  is a gravitational constant. The following results are obtained :

$t$ (seconds) :	1	2	3	4	5
$S$ (feet) :	15	70	140	250	380

Taking  $S$  as the dependent variable, fit a straight line to the data by the method of least squares in a manner that you can estimate  $g$ . What is the estimate of  $g$  ?

5. (a) Explain the method of fitting a second degree parabola by using the principle of least squares.

(b) Fit a parabola  $Y = a + bx + cx^2$  to the following data :

$X :$	1	2	3	4	5	6	7
$Y :$	2.3	5.2	9.7	16.5	29.4	35.5	54.4

6. Fit a second degree parabola to the following data taking  $X$  as the independent variable :

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

Ans.  $Y = -1 + 3.55X - 0.27X^2$

7. In a spectroscopic method for determining the per cent  $X$  of natural rubber content of vulcanizates, the variable  $Y$  used is  $1 + \log_{10} r$ , where  $r$  is the ratio of transmission at two selected wavelengths. In order to establish a relationship between  $X$  and  $Y$ , the following data were obtained :

$X :$	0	20	40	60	80	100
$Y :$	2.19	2.65	3.16	3.57	3.93	4.27

Using least square method, fit a parabola. Comment on your results.

8. Fit a second degree curve  $Y = a + bx + cx^2$  to the following data relating to profit of a certain company.

$Year :$	1980	1982	1984	1986	1988
$Profit$ in lakhs of rupees :	125	140	165	195	230

Estimate the profit in the year 1995.

Ans.  $Y = 114 + 7.2X + 3.15X^2$

9. Explain the method of least squares of fitting a curve to the given mass of data :

$X :$	-2	-1	0	1	2
-------	----	----	---	---	---

$$Y : \quad y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5$$

Fit a parabola  $Y = a + bX + c(X^2 - 2)$ , by the method of least squares and show that

$$a = \bar{y}, \quad b = \frac{1}{10}(-2y_1 - y_2 + y_4 + 2y_5), \quad c = \frac{1}{14}(2y_1 - y_2 + y_4 + 2y_5)$$

10. Show that the best fitting linear function for the points  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  may be expressed in the form

$$\left| \begin{array}{ccc} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{array} \right| = 0, \quad (i = 1, 2, \dots, n).$$

Show that the line passes through the mean point  $(\bar{x}, \bar{y})$ .

### 9.2. Most Plausible Solution of a System of Linear Equations.

Method of least squares is helpful in finding the most plausible values of the variables satisfying a system of independent linear equations whose number is more than the number of variables under study. Consider the following set of  $m$  equations in  $n$  variables  $X, Y, Z, \dots, T$ :

$$\left. \begin{array}{l} a_1 X + b_1 Y + c_1 Z + \dots + k_1 T = l_1 \\ a_2 X + b_2 Y + c_2 Z + \dots + k_2 T = l_2 \\ \vdots \\ a_m X + b_m Y + c_m Z + \dots + k_m T = l_m \end{array} \right\} \quad \dots(9.8)$$

where  $a_i, b_i, \dots, l_i ; i = 1, 2, \dots, m$  are constants.

If  $m = n$ , the system of equations (9.8) can be solved uniquely with the help of algebra. If  $m > n$ , it is not possible to determine a unique solution  $X, Y, Z, \dots, T$  which will satisfy the system (9.8). In this case we find the values of  $X, Y, Z, \dots, T$  which will satisfy the system (9.8) as nearly as possible.

Legendre's principle of least squares consists in minimising the sum of the squares of the 'residuals' or the 'errors'. If

$$E_i = a_i X + b_i Y + c_i Z + \dots + k_i T - l_i ; \quad i = 1, 2, \dots, m$$

is the residual for the  $i$ th equation, then we have to determine  $X, Y, Z, \dots, T$  so that

$$U = \sum_{i=1}^m E_i^2 = \sum_{i=1}^m (a_i X + b_i Y + c_i Z + \dots + k_i T - l_i)^2$$

is minimum.

Using the principle of maxima and minima in differential calculus, the partial derivatives of ' $U$ ' w.r.t.  $X, Y, Z, \dots, T$  should vanish separately. Thus

$$\left. \begin{array}{l} \frac{\partial U}{\partial X} = 0 = \sum_{i=1}^m a_i (a_i X + b_i Y + c_i Z + \dots + k_i T - l_i) \\ \frac{\partial U}{\partial Y} = 0 = \sum_{i=1}^m b_i (a_i X + b_i Y + c_i Z + \dots + k_i T - l_i) \\ \vdots \quad \vdots \\ \frac{\partial U}{\partial T} = 0 = \sum_{i=1}^m k_i (a_i X + b_i Y + c_i Z + \dots + k_i T - l_i) \end{array} \right\} \dots(9.9)$$

These are known as the *normal equations* for  $X, Y, Z, \dots, T$  respectively. Thus we have  $n$  - normal equations in  $n$  unknowns  $X, Y, Z, \dots, T$  and their unique solution gives the best or the most plausible solution of the system (9.8).

Here we see that the *normal equation for any variable is obtained by multiplying each equation by the coefficient of the variable in that equation and then adding all the resulting equations.*

**Example 9.4.** Find the most plausible values of  $X$  and  $Y$  from the following equations :

$$\begin{aligned} X - 5Y + 4 &= 0, & 2X - 3Y + 5 &= 0 \\ X + 2Y - 3 &= 0, & 4X + 3Y + 1 &= 0 \end{aligned}$$

**Solution.** Normal equation for  $X$  is

$$\begin{aligned} 1 \cdot (X - 5Y + 4) + 2 \cdot (2X - 3Y + 5) + 1 \cdot (X + 2Y - 3) + 4 \cdot (4X + 3Y + 1) &= 0 \\ \Rightarrow \quad 22X + 3Y + 15 &= 0 \end{aligned} \dots(*)$$

Normal equation for  $Y$  is

$$\begin{aligned} -5(X - 5Y + 4) - 3(2X - 3Y + 5) + 2(X + 2Y - 3) + 3(4X + 3Y + 1) &= 0 \\ \Rightarrow \quad 3X + 47Y - 38 &= 0 \end{aligned} \dots(**)$$

Solving (\*) and (\*\*), we get  $X = -0.799$  and  $Y = 0.86$ .

Hence the most plausible values of  $X$  and  $Y$  are  $X = -0.80$  (approx.) and  $Y = 0.86$  (approx.)

### EXERCISE 9 (b)

1. Find the most plausible values of  $X$  and  $Y$  from the following equations:

$$(i) \quad \begin{aligned} X + Y &= 3.01, & 2X - Y &= 0.03, \\ .X + 3Y &= 7.03, & 3X + Y &= 4.97. \end{aligned}$$

**Ans.**  $X = 1.0003$ ,  $Y = 2.0007$ .

$$(ii) \quad \begin{aligned} X + Y &= 3, & X - Y &= 2, \\ X + 2Y - 4 &= 0, & X &= 2Y + 1. \end{aligned}$$

2. Find the most plausible values of  $X, Y$  and  $Z$  from the following equations :

$$X - Y + 2Z = 3, \quad 3X + 2Y - 5Z = 5, \quad 4X + Y + 4Z = 21 \text{ and } -X + 3Y + 3Z = 14$$

**Ans.**  $X = 2.47$ ,  $Y = 3.55$ ,  $Z = 1.92$ .

**9.3. Conversion of Data to Linear Form.** Sometimes it may happen that the original data is not in a linear form but can be reduced to linear form by

some simple transformation of variables. We will illustrate this by considering the following curves :

**(a) Fitting of a Power Curve.**  $Y = aX^b$  ... (9-10)

to a set of  $n$  points.

Taking logarithm of both sides, we get

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

$$\Rightarrow U = A + bV$$

where  $U = \log Y$ ,  $A = \log a$  and  $V = \log X$ .

This is a linear equation in  $V$  and  $U$ .

Normal equations for estimating  $A$  and  $B$  are

$$\Sigma U = nA + b\Sigma V \quad \text{and} \quad \Sigma UV = A\Sigma V + b\Sigma V^2 \quad \dots (9-10a)$$

These equations can be solved for  $A$  and  $b$  and consequently, we get

$$a = \text{antilog } (A)$$

With the values of  $a$  and  $b$  so obtained, (9-10) is the curve of best fit to the set of  $n$  points.

**(b) Fitting of Exponential Curves.** (i)  $Y = ab^X$ , (ii)  $Y = ae^{bX}$  to a set of  $n$  points.

$$(i) \quad Y = ab^X \quad \dots (9-11)$$

Taking logarithm of both sides, we get

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

$$\Rightarrow U = A + BX$$

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

This is linear equation in  $X$  and  $U$ .

The normal equations for estimating  $A$  and  $B$  are

$$\Sigma U = nA + B\Sigma X \quad \text{and} \quad \Sigma UX = A\Sigma X + B\Sigma X^2 \quad \dots (9-11a)$$

Solving these equations for  $A$  and  $B$ , we finally get

$$a = \text{antilog } (A) \quad \text{and} \quad b = \text{antilog } (B)$$

With these values of  $a$  and  $b$ , (9-11) is the curve of best fit to the given set of  $n$  points.

$$(ii) \quad Y = ae^{bX} \quad \dots (9-12)$$

$$\log Y = \log a + bX \log e = \log a + (b \log e) X$$

$$\Rightarrow U = A + BX$$

where  $U = \log Y$ ,  $A = \log a$  and  $B = b \log e$ .

This is linear equation in  $X$  and  $U$ .

Thus the normal equations are

$$\Sigma U = nA + B\Sigma X \quad \text{and} \quad \Sigma UX = A\Sigma X + B\Sigma X^2 \quad \dots (9-12a)$$

From these we find  $A$  and  $B$  and consequently

$$a = \text{antilog } (A) \quad \text{and} \quad b = \frac{B}{\log e}$$

**Example 9.5.** Fit an exponential curve of the form  $Y = ab^x$  to the following data :

X:	1	2	3	4	5	6	7	8
Y:	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

**Solution.**

	X	Y	$U = \log Y$	XU	$X^2$
	1	1.0	0.0000	0.0000	1
	2	1.2	0.0792	0.1584	4
	3	1.8	0.2553	0.7659	9
	4	2.5	0.3979	1.5916	16
	5	3.6	0.5563	2.7815	25
	6	4.7	0.6721	4.0326	36
	7	6.6	0.8195	5.7365	49
	8	9.1	0.9590	7.6720	64
Total	36	30.5	3.7393	22.7385	204

(9.11a) gives the normal equations as

$$3.7393 = 8A + 36B$$

$$\text{and} \quad 22.7385 = 36A + 204B$$

Solving, we get

$$B = 0.1408 \text{ and } A = -0.1662 = \bar{T} \cdot 8338$$

$$\therefore b = \text{Antilog } B = 1.383 \text{ and } a = \text{Antilog } A = 0.6821$$

Hence the equation of the required curve is

$$Y = 0.6821(1.38)^X$$

**Example 9.6.** Derive the least square equations for fitting a curve of the type  $Y = aX + (b/X)$ , to a set of  $n$  points  $(x_i, y_i)$ ;  $i = 1, 2, \dots, n$ .

**Solution.** The error of estimate  $E_i$  for the  $i$ th point  $(x_i, y_i)$  is given by

$$E_i = \left( y_i - ax_i - \frac{b}{x_i} \right)$$

According to the principle of least squares, we have to determine the values of  $a$  and  $b$  so that sum of the squares of errors  $E$ , viz.,

$$E = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n \left( y_i - ax_i - \frac{b}{x_i} \right)^2$$

is minimum.

Consequently, the normal equations are

$$\frac{\partial E}{\partial a} = 0 = -2 \sum_{i=1}^n x_i \left( y_i - ax_i - \frac{b}{x_i} \right)$$

$$\frac{\partial E}{\partial b} = 0 = -2 \sum_{i=1}^n \frac{1}{x_i} \left( y_i - ax_i - \frac{b}{x_i} \right)$$

which on simplification give

$$\sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i^2 + nb$$

$$\text{and } \sum_{i=1}^n \left( \frac{y_i}{x_i} \right) = na + b \sum_{i=1}^n \left( \frac{1}{x_i^2} \right)$$

**Example 9.7.** Three independent measurements on each of the three angles A, B, C of a triangle are as follows :

A	B	C
39.5	60.3	80.1
39.3	62.2	80.3
39.6	60.1	80.4

Obtain the best estimates of the three angles taking into account the relation that the sum of the angles is equal to  $180^\circ$ .

**Solution.** Let the three observations on A be denoted by  $x_1, x_2, x_3$ , on B by  $y_1, y_2, y_3$  and on C by  $z_1, z_2, z_3$ . Let  $\theta_1, \theta_2$  be the best estimates for A and B respectively.

According to the principle of least squares, our problem is to estimate  $\theta_1$  and  $\theta_2$ , so that

$$E = \sum (x_i - \theta_1)^2 + \sum (y_i - \theta_2)^2 + \sum (z_i - 180 + \theta_1 + \theta_2)^2$$

is minimum, summation being taken over i from 1 to 3.

Equating to zero the partial derivatives of E w.r.t.  $\theta_1$  and  $\theta_2$ , the normal equations are

$$\frac{\partial E}{\partial \theta_1} = 0 = -\sum (x_i - \theta_1) + \sum (z_i - 180 + \theta_1 + \theta_2) \quad ...(*)$$

$$\frac{\partial E}{\partial \theta_2} = 0 = -\sum (y_i - \theta_2) + \sum (z_i - 180 + \theta_1 + \theta_2) \quad ...(**)$$

From (\*) and (\*\*), we get

$$\begin{aligned} 3\theta_1 - \sum x_i + \sum z_i - 540 + 3\theta_1 + 3\theta_2 &= 0 \\ 3\theta_2 - \sum y_i + \sum z_i - 540 + 3\theta_1 + 3\theta_2 &= 0 \end{aligned} \quad ...(***)$$

$$\text{But } \sum x_i = 39.5 + 39.3 + 39.6 = 118.4$$

$$\sum y_i = 60.3 + 62.2 + 60.1 = 182.6$$

$$\sum z_i = 80.1 + 80.3 + 80.4 = 240.8$$

Substituting in (\*\*), we get

$$6\theta_1 + 3\theta_2 - 417.6 = 0 \quad \text{and} \quad 3\theta_1 + 6\theta_2 - 481.8 = 0$$

$$\therefore \hat{A} = \hat{\theta}_1 = 39.27, \hat{B} = \hat{\theta}_2 = 60.66 \quad \text{and} \quad \hat{C} = 180 - \hat{\theta}_1 - \hat{\theta}_2 = 80.07$$

**9·4, Selection of Type of Curve to be Fitted.** The greatest limitation of the method of curve fitting by the principle of least squares is the choice of the mathematical curve to be fitted to the given data. The choice of a particular curve for describing the given data requires great skill, intelligence and expertise. The graph of the given data enables us to have a fairly good idea about the type of the curve to be fitted. The graph will clearly reveal if the trend is linear (straight line) or curvilinear (non-linear). If the graph exhibits a curvilinear trend then further approximations to the type of trend curve can be obtained on plotting the data on a semi-logarithmic scale. A careful study of the graph obtained on plotting the data on an arithmetic or semi-logarithmic scale often provides adequate basis for selecting the type of the curve. The various types of curves that may be used to describe the given data in practise are: [If  $y_x$  is the value of the dependent variable corresponding to the value  $x$  of the independent variable]

- (i) *A straight line:*  $y_x = a + bx$
- (ii) *Second degree parabola:*  $y_x = a + bx + cx^2$
- (iii) *kth degree polynomial :*  $y_x = a_0 + a_1 x + a_2 x^2 + \dots + a_k x^k$
- (iv) *Exponential curve:*  $y_x = ab^x$

$$\Rightarrow \log y_x = \log a + x \log b = A + Bx, \text{ (say).}$$

- (iv) *Second degree curve fitted to logarithms:*

$$\begin{aligned} & y_x = ab^x \quad cx^2 \\ \Rightarrow & \log y_x = \log a + x \log b = x^2 \log c \\ & \qquad \qquad \qquad = A + Bx + Cx^2, \text{ (say).} \end{aligned}$$

- (vi) *Growth Curves:*

$$(a) \quad y_x = a + b^x \quad (\text{Modified Exponential Curve})$$

$$(b) \quad y_x = abc^x \quad (\text{Gompertz Curve})$$

$$\Rightarrow \log y_x = \log a + c^x \cdot \log b = A + Bc^x, \text{ (say)}$$

$$(c) \quad y_x = \frac{k}{1 + \exp(a + bx)}, \quad b < 0 \quad (\text{Logistic Curve})$$

For decideing about the type of curve to be fitted to a given set of data, the following points may be helpful:

(i) When the  $y_x$  series is found to be increasing by equal absolute amounts, the straight line curve is used. In this case, the graph of the data will give a straight line graph.

(ii) The logarithmic straight line (exponential curve  $y_x = ab^x$ ) is used when the series is increasing or decreasing by a constant percentage rather than a constant absolute amount. In this case, the data plotted on a semi-logarithmic scale will give a straight line graph.

(iii) Second degree curve fitted to logarithms may be tried if the data plotted

on a semi-logarithmic scale is not a straight line graph but shows curvature, being concave either upward or downward.

For further guidelines, the following statistical tests based on the calculus of finite differences [c.f. Chapter 17] may be applied.

We know that for a polynomial  $y_x$  of  $n$ th degree in  $x$ ,

$$\Delta^r y_x = \text{constant}, r = n \\ = 0 \quad , \quad r > n \}$$

where  $\Delta$  is the difference operator given by  $\Delta y_x = y_{x+h} - y_x$ ,  $h$  being the interval of differencing and  $\Delta^r y_x$  is the  $r$ th order difference of  $y_x$ .

1. If  $\Delta y_x = \text{constant}$ , use straight line curve.
2. If  $\Delta^2 y_x = \text{constant}$ , use a second degree (parabolic) curve.
3. If  $\Delta (\log y_x) = \text{constant}$ , use exponential curve.
4. If  $\Delta^2 (\log y_x) = \text{constant}$ , use second degree curve fitted to logarithms.
5. If  $\Delta y_x$  tends to decrease by a constant percentage, use modified exponential curve.
6. If  $\Delta y_x$  resembles a skewed frequency curve, use a Gompertz curve or Logistic curve.
7. The growth curves, viz., modified exponential, Gompertz and Logistic curves, can be approximated by the constancy of the ratios

$$\frac{\Delta y_x}{\Delta y_{x-1}}, \left\{ \frac{\Delta \log y_x}{\Delta \log y_{x-1}} \right\}, \left\{ \frac{\Delta(1/y_x)}{\Delta(1/y_{x-1})} \right\},$$

respectively for all possible values of  $x$ .

### EXERCISE 9 (c)

1. Describe the method of fitting the following curves :

$$(i) Y = ae^{bx}, \quad (ii) Y = aX^b$$

2. (a) Fit an equation of the form  $Y = ab^X$  to the following data :

X:	2	3	4	5	6
Y:	144	172.8	207.4	248.8	298.6

Ans.  $Y = (101.3)(1.196)^X$

- (b) Fit a curve of the type  $Y = ab^X$  to the following data :

X:	2	3	4	5	6
Y:	8.3	15.4	33.1	65.2	127.4

Estimate  $Y$  when  $X = 4.5, 7$  and  $3.5$ .

3. Fit a curve of the form  $Y = bc^X$  to the following data :

Year (X):	1951	1952	1953	1954	1955	1956	1957
Production							

in tons (Y):	201	263	314	395	427	504	612
--------------	-----	-----	-----	-----	-----	-----	-----

4. In an experiment in which the growth of duck weed under certain con-

ditions was measured, the following results were obtained :

Weeks ( $X$ ) ...	0	1	2	3	4	2	6	7	8
No. of friends ( $Y$ ) ...	20	30	52	77	135	211	326	550	1052

Assuming the relationship of the form  $Y = ae^{bx}$ , find the best values of  $a$  and  $b$  by the method of least squares.

5. For the data given below, find the equation to the best fitting exponential curve of the form  $Y = ae^{bx}$ .

$X$ :	1	2	3	4	5	6
$Y$ :	1.6	4.5	13.8	40.2	125.0	300.0

$$\text{Ans. } Y = (0.557) e^{1.05X}$$

6. Fit the curve  $Y = aX^2 + (b/X)$  to the following data :

$X$ :	1	2	3	4
$Y$ :	-1.51	0.99	3.88	7.66

7. The following table gives corresponding values of two variables  $X$  and  $Y$ .

$X$ :	1	2	3	4	5
$Y$ :	1.8	5.1	8.9	14.1	19.8

It is found that they are connected by a law of the form  $Y = aX + bX^2$ , where  $a$  and  $b$  are constants. Find the best values of  $a$  and  $b$  by the method of least squares. Calculate the value of  $Y$  for  $X = 2$ .

$$\text{Ans. } a = 1.521; b = 0.49; 5.006$$

8. The following pairs of observations were noted in experimental work on cosmic rays. Find, by the method of least squares, the best values of  $a$  and  $b$  for the equation  $\log R = a - bC$  which fits the data and estimate the most probable value of  $R$  for  $C = 20.7$ .

$C$ :	14	15	16	17	18
$R$ :	24.1	20.5	14.0	7.3	5.0

9. (a) Explain the principle of least squares and describe its applications in fitting a curve of the form  $Y = a \exp(bX + cX^2)$ .

(b) Fit an indifference curve of the type  $XY = b + aX$  to the data given below:

Consumption of Commodity X :	1	2	3	4
Consumption of Commodity Y :	3	1.5	6	7.5

Hint.  $y = a + (b/x)$ . Now proceed as in Example 9.6.

$$\text{Ans. } XY = 1.3X + 1.7$$

10. (a) Show that the parabola of best fit for the points

$$(x_1, y_1); (x_2, y_2); \dots; (x_{2n+1}, y_{2n+1})$$

where the values of  $x$  are in A.P. with common difference unity and  $\bar{x} = 0$ , can be expressed in the form

$$\left| \begin{array}{cccc} y & \frac{n(n+1)(2n+1)}{3} & \bar{x} & \frac{1}{an+1} \\ \sum y_i & 0 & 0 & 0 \\ \sum x_i y_i & \frac{n^2(n+1)^2}{2} & \frac{n(n+1)(2n+1)}{3} & \frac{n(n+1)(2n+1)}{3} \end{array} \right| = 0$$

[Delhi Univ. B.A. (Pass), 1984]

**Hint.** Use (9.4b), with

$$x_i = a + i ; \quad i = 1, 2, \dots, (2n+1). \text{ Since } \bar{x} = 0,$$

$$\sum x_i = (2n+1)a + \sum i \Rightarrow 0 = (2n+1)a + \frac{(2n+1)(2n+2)}{2}$$

$$\Rightarrow a = -(n+1)$$

$$\sum x_i^2 = \sum (a+i)^2 = (2n+1)a^2 + \sum i^2 + 2a \sum i$$

and so on, for  $\sum x_i^3$  and  $\sum x_i^4$ .

$$\left[ \sum_{i=1}^m i^2 = \frac{m(m+1)(2m+1)}{6}; \quad \sum_{i=1}^m i^3 = \left[ \frac{m(m+1)}{2} \right]^2 \right]$$

$$\text{and } \sum_{i=1}^m i^4 = \frac{1}{30} m(m+1)(2m+1)(3m^2+3m-1)$$

(b) When do we prefer logarithmic curve to ordinary curve?

**9.5. Curve Fitting by Orthogonal Polynomials.** Suppose that the polynomial of  $p$ th degree of  $Y$  on  $X$  is

$$Y = a_0 + a_1 X + a_2 X^2 + \dots + a_p X^p \quad \dots(9.13)$$

The normal equations for determining the constants  $a_i$ 's are obtained by the principle of least squares by minimising the residual or error sum of squares

$$E = \sum (y - a_0 - a_1 x - a_2 x^2 - \dots - a_p x^p)^2 \quad \dots(9.14)$$

summation being extended over the given set of observations. The normal equations are :

$$\frac{\partial E}{\partial a_j} = 0, \quad (j = 0, 1, 2, \dots, p)$$

$$\text{i.e., } \sum x^j (y - a_0 - a_1 x - a_2 x^2 - \dots - a_p x^p) = 0, \quad [j = 0, 1, 2, \dots, p] \quad \dots(9.15)$$

Assume that  $X$  and  $Y$  are measured from their means (and this we can do without any loss of generality) so that

$$\mu_r = \mu'_r = E(X') = \frac{1}{N} \sum x'$$

and write,

$$\mu_{j1} = \frac{1}{N} \sum x^{j1} \cdot y,$$

where  $N$  is number of observations taken on each of the variables  $X$  and  $Y$ . Hence (9.15) gives

$$\begin{aligned} \mu_j - a_0 \mu_j - a_1 \mu_{j+1} - a_2 \mu_{j+2} - \dots - a_p \mu_{j+p} &= 0 ; \quad j = 0, 1, 2, \dots, p \\ \Rightarrow a_0 \mu_j + a_1 \mu_{j+1} + a_2 \mu_{j+2} + \dots + a_p \mu_{j+p} &= \mu_{j1} ; \quad j = 0, 1, 2, \dots, p \end{aligned}$$

Putting  $j = 0, 1, 2, \dots, p$ , we get respectively

$$\left. \begin{array}{l} a_0 \mu_0 + a_1 \mu_1 + a_2 \mu_2 + \dots + a_p \mu_p = \mu_{01} \\ a_0 \mu_1 + a_1 \mu_2 + a_2 \mu_3 + \dots + a_p \mu_{p+1} = \mu_{11} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ a_0 \mu_p + a_1 \mu_{p+1} + a_2 \mu_{p+2} + \dots + a_p \mu_{2p} = \mu_{p1} \end{array} \right\} \quad \dots(9.16)$$

Solving (9.16) for  $a_0, a_1, \dots, a_p$  in terms of the moments  $\mu_j$ 's and  $\mu_{ji}$ 's,  $j = 0, 1, 2, \dots, p$  and substituting in (9.13) we get the required curve of best fit.

Let

$$\Delta^{(p)} = \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \dots & \mu_p \\ \mu_1 & \mu_2 & \mu_3 & \dots & \mu_{p+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_p & \mu_{p+1} & \mu_{p+2} & \dots & \mu_{2p} \end{vmatrix} \quad \dots(9.17)$$

and  $\Delta_j^{(p)}$  be the determinant obtained on replacing  $(j+1)$ th column of  $\Delta^{(p)}$  by the column

$$\begin{pmatrix} \mu_{01} \\ \mu_{11} \\ \vdots \\ \mu_{p1} \end{pmatrix} \quad \text{then} \quad a_j = \frac{\Delta_j^{(p)}}{\Delta^{(p)}} \quad \dots(9.18)$$

The required curve of best fit is the eliminant of  $a_i$ 's in (9.13) and (9.16) and is given by

$$\begin{vmatrix} Y & 1 & X & X^2 & \dots & X^p \\ \mu_{01} & \mu_0 & \mu_1 & \mu_2 & \dots & \mu_p \\ \mu_{11} & \mu_1 & \mu_2 & \mu_3 & \dots & \mu_{p+1} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \mu_{p1} & \mu_p & \mu_{p+1} & \mu_{p+2} & \dots & \mu_{2p} \end{vmatrix} = 0 \quad \dots(9.19)$$

The use of equation (9.19) is subject to one serious drawback. If we have a set of data and apart from inspection if there is no guide regarding the order of the polynomial to be fitted, the only way left to us is to try curves of order 1, 2, 3, ... until we reach the point where further terms do not improve the fit. Every time we add a new term, the  $a_i$ 's given by (9.18) change and accordingly the determinantal arithmetic has to be done afresh. For example, if we want to fit a polynomial curve of third or higher degree to the same data then we cannot use the coefficients which we computed while fitting a second degree parabola. To overcome this drawback Prof. R.A. Fisher suggested a method which involved the fitting of *Orthogonal Polynomials* by the principle of least squares, so that each term is independent of the other, i.e., each of the coefficients in

the polynomial is independent of the other so that each of them can be calculated independently. In this method, the coefficients computed earlier remain the same and we have to compute the coefficient only for the added term.

**9.5.1. Orthogonal Polynomials (Def.).** Two polynomials  $P_1(x)$  and  $P_2(x)$  are said to be *orthogonal* to each other if

$$\sum P_1(x) P_2(x) = 0, \quad \dots(9.20)$$

where summation is taken over a specified set of values of  $x$ . If  $x$  were a continuous variable in the range from  $a$  to  $b$ , the condition for orthogonality gives

$$\int_a^b P_1(x) P_2(x) dx = 0 \quad \dots(9.20a)$$

For example, if we take

$$P_0 = 1, P_1(x) = x - 4, P_2(x) = x^2 - 8x + 12, P_3(x) = x^3 - 12x^2 + 41x - 36 \quad \dots(9.20b)$$

then these are orthogonal to each other for a set of integral values of  $x$  from 1 to 7 as explained in the following table. Other examples of orthogonal polynomials are Hermite polynomials, Gram Charlier's polynomials, Legender's polynomials, etc.

### ORTHOGONALITY OF POLYNOMIALS DEFINED IN (9.20b)

$x$	$P_0 P_1$	$P_0 P_2$	$P_0 P_3$	$P_1 P_2$	$P_1 P_3$	$P_2 P_3$
1	-3	5	-6	-15	18	-30
2	-2	0	6	0	-12	0
3	-1	-3	6	3	-6	-18
4	0	-4	0	0	0	0
5	1	-3	-6	-3	-6	18
6	2	0	-6	0	-12	0
7	3	5	6	15	18	30
Total	0	0	0	0	0	0

**9.5.2. Fitting of Orthogonal Polynomials.** The  $p$ th degree polynomial (9.13) can be rewritten as

$$Y = b_0 P_0 + b_1 P_1 + b_2 P_2 + \dots + b_p P_p \quad \dots(9.21)$$

where  $P_j$ 's are polynomials in  $x$ ,  $P_j$  being a polynomial of degree  $j$ , ( $j = 0, 1, 2, \dots, p$ ). We shall determine  $P_j$ 's so that they satisfy the condition of orthogonality, viz.,

$$\sum_x P_j(x) P_k(x) = 0; j \neq k \quad \dots(9.22)$$

the summation being extended over the observed values of  $x$ . The normal equations for estimating the constants  $b_j$ 's are obtained on minimising

$$E = \sum (y - b_0 P_0 - b_1 P_1 - \dots - b_p P_p)^2 \quad \dots(9.23)$$

and are given by

$$\frac{\partial E}{\partial b_j} = 0$$

$$\Rightarrow \sum P_j (y - b_0 P_0 - b_1 P_1 - \dots - b_p P_p) = 0 : j = 0, 1, 2, \dots, p.$$

Simplifying and using (9.22), we get

$$\begin{aligned} & \sum P_j \cdot y - b_j \sum P_j^2 = 0 \\ \Rightarrow \quad & b_j = \frac{\sum y P_j}{\sum P_j^2}, j = 0, 1, 2, \dots, p. \end{aligned} \quad \dots(9.24)$$

Thus  $b_j$  is determined by  $P_j$ . If having fitted a curve of order  $p$  we wish to go a step further by adding a term  $b_{p+1} P_{p+1}$ , the coefficients already obtained in (9.24) remain unaltered.

Moreover, the use of orthogonal polynomials will give us a very convenient method of determining, step by step, the goodness of fit of the polynomial curve. For  $p$ th degree polynomial (9.21), the error sum of squares is [c.f. (9.23)]

$$\begin{aligned} E &= \sum (y - b_0 P_0 - b_1 P_1 - \dots - b_p P_p)^2 \\ &= \sum y^2 + b_0^2 \sum P_0^2 + b_1^2 \sum P_1^2 + \dots + b_p^2 \sum P_p^2 \\ &\quad - 2 b_0 \sum y P_0 - 2 b_1 \sum y P_1 - \dots - 2 b_p \sum y P_p, \end{aligned}$$

other terms vanish because of orthogonality conditions (9.22). Using (9.24) we finally obtain

$$E = \sum y^2 - b_0^2 \sum P_0^2 - b_1^2 \sum P_1^2 - \dots - b_p^2 \sum P_p^2 \quad \dots(9.25)$$

Thus the effect of adding any term  $b_j P_j$  is to reduce the error (residual) sum of squares  $E$  by  $b_j^2 \sum P_j^2$  and we may examine the effect of this term on  $E$  separately. If we find that the addition of any term  $b_p P_p$  does not reduce  $E$  significantly, we may conclude that it is not desired (as far as the representation of the given data by a polynomial curve is concerned).

### 9.5.3. Finding The Orthogonal Polynomial $P_p$ , in Fitting a Polynomial of Degree $p$ .

Let  $P_p$ , the polynomial of degree  $p$  in  $x$  be given by

$$P_p = \sum_{j=0}^p c_{pj} x^j \quad \dots(9.26)$$

This contains  $(p+1)$  unknown constants  $c_{p0}, c_{p1}, \dots, c_{pp}$ . Hence in all the polynomials in (9.21) up to and including those of  $p$ th order, there are

$$1 + 2 + 3 + \dots + (p+1) = \frac{(p+1)(p+2)}{2},$$

unknown constants. The orthogonality conditions

$$\sum P_i P_j = 0, i \neq j = 0, 1, 2, \dots, p,$$

provide  $p+1$  conditions on the  $c$ 's so that there are

$$\frac{(p+1)(p+2)}{2} - \frac{(p+1)p}{2} = p+1,$$

constants which can be assigned arbitrarily. We will take one for each polynomial  $P_j$  ( $j = 0, 1, 2, \dots, p$ ) and assign it such that the coefficient of  $x^j$  in  $P_j$  is unity i.e.,

$$c_{jj} = 1, j = 0, 1, 2, \dots, p, \quad \dots(9.27)$$

In particular  $c_{00} = P_0 = 1$ . The orthogonality conditions give:

$$\sum_x P_p P_j = 0, j < p \quad (j = 0, 1, 2, \dots, p-1) \quad \dots(9.28)$$

$$j = 0, \text{ gives } \sum P_p P_0 = 0 \Rightarrow \sum P_p = 0; (\because P_0 = 1) \quad \dots(*)$$

$$j = 1, \text{ gives } \sum P_p P_1 = 0 \Rightarrow \sum P_p = 0, (x + k) = 0$$

$$\Rightarrow \sum P_p \cdot x + k \sum P_p = 0$$

$$\Rightarrow \sum x P_p = 0 \quad \dots(**)$$

$$j = 2, \text{ gives } \sum P_p P_2 = 0 \Rightarrow \sum P_p = (x^2 + k_1 x + k_2) = 0 \quad [\text{Using } (*)]$$

$$\Rightarrow \sum x^2 P_p = 0 \quad [\text{Using } (*) \text{ and } (**)]$$

Similarly proceeding, we shall get in general

$$\sum_x P_p x^r = 0, \quad r = 0, 1, 2, \dots, p-1 \quad \dots(9.29)$$

$$\Rightarrow \sum_x \left( \sum_{j=0}^p c_{pj} \cdot x^j \right) x^r = 0$$

$$\Rightarrow \sum_{j=0}^p \left( c_{pj} \sum_x x^{j+r} \right) = 0$$

Dividing both sides by  $N$ , the number of observations on each of the variables  $X$  and  $Y$ , we get.

$$\sum_{j=0}^p c_{pj} \mu_{j+r} = 0; \quad r = 0, 1, 2, \dots, (p-1) \quad \dots(9.30)$$

where  $x$  is assumed to be measured from mean. Putting  $r = 0, 1, 2, \dots, (p-1)$  in (9.30), we get respectively

$$c_{p0} \mu_0 + c_{p1} \mu_1 + \dots + c_{pj} \mu_j + \dots + c_{p,p-1} \mu_{p-1} + c_{pp} \mu_p = 0$$

$$c_{p0} \mu_1 + c_{p1} \mu_2 + \dots + c_{pj} \mu_{j+1} + \dots + c_{p,p-1} \mu_p + c_{pp} \mu_{p+1} = 0$$

$$c_{p0} \mu_{p-1} + c_{p1} \mu_p + \dots + c_{pj} \mu_{j+p-1} + \dots + c_{p,p-1} \mu_{2p-2} + c_{pp} \mu_{2p-1} = 0$$

Noting that  $c_{pp} = 1$ , solving the above equations for  $c$ 's, we get

$$c_{pj} = \frac{\begin{vmatrix} \mu_0 & \mu_1 & \dots & -\mu_p & \dots & \mu_{p-1} \\ \mu_1 & \mu_2 & \dots & -\mu_{p+1} & \dots & \mu_p \\ \vdots & \vdots & & \vdots & & \vdots \\ \mu_{p-1} & \mu_p & \dots & \mu_{2p-1} & \dots & \mu_{2p-2} \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 & \dots & \mu_j & \dots & \mu_{p-1} \\ \mu_1 & \mu_2 & \dots & \mu_{j+1} & \dots & \mu_p \\ \vdots & \vdots & & \vdots & & \vdots \\ \mu_{p-1} & \mu_p & \dots & \mu_{j+p-1} & \dots & \mu_{2p-2} \end{vmatrix}} = \frac{\Delta^{(p)} p j}{\Delta^{(p-1)}} \quad \dots(9.31)$$

where  $\Delta^{(p)}$  has been defined in (9.17) and  $\Delta^{(p)}_{pj}$  is the minor of the element in the last row and  $(j+1)$ th column in  $\Delta^{(p)}$ . Substituting this value of  $c_{pj}$  in (9.26), we get

$$\begin{aligned} P_p &= \sum_{j=0}^p c_{pj} x^j = \sum_{j=0}^p \frac{\Delta^{(p)}_{pj}}{\Delta^{(p-1)}} \cdot x^j \\ &= \frac{1}{\Delta^{(p-1)}} \sum_{j=0}^p \Delta^{(p)}_{pj} x^j \\ &= \frac{1}{\Delta^{(p-1)}} \left[ \Delta^{(p)}_{p0} + x \Delta^{(p)}_{p1} + \dots + x^p \cdot \Delta^{(p)}_{pp} \right] \\ \Rightarrow P_p &= \frac{1}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_p \\ \mu_1 & \mu_2 & \mu_3 & \mu_{p+1} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{p-1} & \mu_p & \mu_{p+1} & \mu_{2p-1} \\ 1 & x & x^2 & x^p \end{vmatrix} \quad \dots(9.32) \end{aligned}$$

In particular if  $\mu_0 = 1, \mu_1 = 0$  and  $\mu_2 = 1$ , i.e., if  $x$  is a standardised variate then the orthogonal polynomials are given by

$$P_0 = 1 \quad \dots(9.33)$$

$$P_1(x) = \frac{\begin{vmatrix} \mu_0 & \mu_1 \\ 1 & x \end{vmatrix}}{\mu_0} = x \quad \dots(9.33a)$$

$$P_2(x) = \frac{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ 1 & x & x^2 \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{vmatrix}} = x^2 - \mu_3 x - 1 \quad \dots(9.33b)$$

$(\because \mu_0 = 1, \mu_1 = 0, \mu_2 = 1)$

$$P_3(x) = \frac{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_3 \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 \\ \mu_2 & \mu_3 & \mu_4 & \mu_5 \\ 1 & x & x^2 & x^3 \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix}} + \frac{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix}} \quad \dots(9.33c)$$

and so on.

If we further assume that  $x$  is a standard *normal* variate so that  $\mu_3 = \mu_5 = \dots = \mu_{2r+1} = 0$ , then the above orthogonal polynomials are called *Hermite Polynomials* and are given by

$$P_0 = 1 ; P_1(x) = x ; P_2(x) = x^2 - 1 ; P_3(x) = x^3 - 3x ; P_4(x) = x^4 - 6x^2 + 3 ;$$

and so on, where  $x$  is a continuous r.v. taking values from  $-\infty$  to  $\infty$ .  $\dots(9.34)$

**Remark.** Hermite Polynomials defined in (9.34) are orthogonal w.r.t. the weight function

$$\alpha(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right); -\infty < x < \infty$$

i.e.,  $\int_{-\infty}^{\infty} P_i(x) P_j(x) \alpha(x) dx = 0 ; i \neq j$  ... (9.35)

where  $P_1(x), P_2(x), P_3(x), P_4(x)$  are defined in (9.34).

**9.5.4. Determination of the Coefficients  $b_j$ 's in (9.21).** From (9.24), we get

$$b_p = \sum y P_p / \sum P_p^2 \quad \dots (9.36)$$

$$\begin{aligned} \text{Now } \sum P_p^2 &= \sum P_p P_p \\ &= \sum_x P_p [c_{p0} + c_{p1}x + c_{p2}x^2 + \dots + c_{pp}x^p] \\ &= \sum_x P_p \cdot x^p, \end{aligned}$$

on using (9.29) and the fact that  $c_{pp} = 1$ .

$$\begin{aligned} \sum P_p^2 &= \sum_x \left( \sum_{j=0}^p c_{pj} x^j \right) x^p = \sum_{j=0}^p \left( c_{pj} \sum_x x^{p+j} \right) \\ &= N \sum_{j=0}^p c_{pj} \mu_{p+j} = N \sum_{j=0}^p \frac{\Delta^{(p)}_{pj}}{\Delta^{(p-1)}} \cdot \mu_{p+j} \quad [\text{From (9.31)}] \\ &= \frac{N}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0 & \mu_1 & \dots & \mu_p \\ \mu_1 & \mu_2 & \dots & \mu_{p+1} \\ \vdots & & & \\ \mu_{p-1} & \mu_p & \dots & \mu_{2p-1} \\ \mu_p & \mu_{p+1} & \dots & \mu_{2p} \end{vmatrix} \end{aligned}$$

[Proceeding exactly as we obtained (9.32)]

$$= \frac{N \Delta^{(p)}}{\Delta^{(p-1)}} \quad \dots (9.37)$$

$$\begin{aligned} \text{Similarly, } \sum y P_p &= N \sum_{j=0}^p \frac{\Delta^{(p)}_{pj}}{\Delta^{(p-1)}} \cdot \mu_{j1} \\ &= \frac{N}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0 & \mu_1 & \dots & \mu_p \\ \mu_1 & \mu_2 & \dots & \mu_{p+1} \\ \vdots & & & \\ \mu_{p-1} & \mu_p & \dots & \mu_{2p-1} \\ \mu_{p1} & \mu_{11} & \dots & \mu_{p1} \end{vmatrix} \\ &= \frac{N \cdot \Delta^{(p)}}{\Delta^{(p-1)}} \quad \dots (9.38) \end{aligned}$$

where  $\Delta^{(p)}$  and  $\Delta^{(p)}_{pj}$  are defined in (9.17). Substituting in (9.36) we get

...(9.39)

If the variable  $x$  takes the integral values 1, 2, ...,  $N$ , then the first seven of these orthogonal polynomials  $P_j$ 's  $j = 0, 1, 2, 3, \dots, 6$  are given by :

$$\begin{aligned} P_0(x) &= 1, P_1(x) = \lambda_1 \cdot \xi \\ P_2(x) &= \lambda_2 \left\{ \xi^2 - \frac{N^2 - 1}{12} \right\} \\ P_3(x) &= \lambda_3 \left\{ \xi^3 - \frac{3N^2 - 7}{20} \xi \right\} \\ P_4(x) &= \lambda_4 \left\{ \xi^4 - \frac{3N^2 - 13}{14} \xi^2 + \frac{3}{560} (N^2 - 1)(N^2 - 9) \right\} \\ P_5(x) &= \lambda_5 \left\{ \xi^5 - \frac{5}{18} (N^2 - 7) \xi^3 + \frac{1}{1008} (15N^4 - 230N^2 + 407) \xi \right\} \\ P_6(x) &= \lambda_6 \left\{ \xi^6 - \frac{5}{44} (3N^2 - 31) \xi^4 + \frac{1}{176} (5N^4 - 110N^2 + 329) \xi^2 \right\} \\ &\quad - \frac{5}{14784} (N^2 - 1)(N^2 - 9)(N^2 - 25) \} \end{aligned}$$

and so on, where  $\xi = x - \bar{x}$  so that  $\sum \xi = 0$  and  $\lambda_i$ 's are arbitrary constants.

If  $y = b_0 + b_1 P_1(x) + b_2 P_2(x) + \dots + b_p P_p(x)$ ; is the orthogonal polynomial fitted to the given data then, using (9.24), we get

$$\left. \begin{aligned} b_0 &= \frac{\sum y P_0}{\sum P_0^2} = \frac{\sum y}{N}; (\because P_0 = 1) \\ b_i &= \frac{\sum_x y P_i}{\sum_x P_i^2}, (i = 1, 2, \dots, p) \end{aligned} \right\} \quad \dots(9.40)$$

The origin of  $P_i$ 's is so chosen that  $\sum P_i = 0$ .

If  $N$ , the number of observations is odd, then we take

$$\xi = \frac{x_j - A}{h}$$

and if  $N$  is even then we take

$$\xi = \frac{x_i - A_1}{(h/2)}$$

where  $h$  = length of the interval (for values of  $x$ )

$A$  = middle value (item) of the data

and  $A_1$  = Arithmetic mean of two middle values of the data.

The values of  $P_i$ 's and  $\lambda_i$ 's are obtained from 'Statistical Tables' by

R.A. Fisher for the values of N from 3 to 75. In these tables the orthogonal polynomials  $P_i$ 's are denoted by  $\phi_i$ 's. We reproduce below these tables for  $N = 3$  to  $N = 6$ .

### TABLES OF ORTHOGONAL POLYNOMIALS

$N = 3$			$N = 4$			$N = 5$			
$\phi_1$	$\phi_2$	$\phi_3$	$\phi_1$	$\phi_2$	$\phi_3$	$\phi_1$	$\phi_2$	$\phi_3$	$\phi_4$
-1	1	-3	1	-1	-2	2	-1	1	
0	-2	-1	-1	-3	-1	-1	2	-4	
1	1	1	-1	-3	0	-2	0	6	
		3	1	1	1	-1	-2	-4	
					2	2	1	1	
$\sum_x \phi_i^2$	2	6	20	4	20	10	14	10	70
$\lambda_i$	1	3	2	1	3	1	1	$\frac{5}{6}$	$\frac{35}{22}$
$N = 6$									
$\phi_1$	$\phi_2$	$\phi_3$	$\phi_4$	$\phi_5$					
-5	5	-5	1	-1					
-3	-1	7	-3	5					
-1	-4	4	2	-10					
1	-4	-4	-2	10					
3	-1	-7	-3	-5					
5	5	5	1	1					
$\sum_x \phi_i^2$	70	84	180	28	252				
$\lambda_i$	2	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{7}{12}$	$\frac{21}{10}$				

**Example 9.8.** Fit a straight line  $y = a + bx\dots(*)$  to the following data by using orthogonal polynomials.

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

**Solution.** Here  $N = 5$ . Let us transform to the variable

$$\xi = \frac{x-2}{1} = x-2 \text{ so that } \sum \xi = 0$$

Let the orthogonal polynomial form of straight line (\*) be

$$y = b_0 + b_1 P_1(x) = b_0 + b_1 \phi_1(x) \quad .(**) \quad$$

$x$	$\xi = x - 2$	$y$	$\phi_1$	$y \phi_1$	
0	-2	1	-2	-2	
1	-1	1.8	-1	-1.8	
2	0	3.3	0	0	
3	1	4.5	1	4.5	
4	2	6.3	2	12.6	
Total		16.9	0	13.3	

The values of  $\phi_1$  are noted from the tables for  $N = 5$ . From tables we also find

$$\sum \phi_1^2 = 10, \lambda_1 = 1$$

$$\text{Now using (9.40), } b_0 = \frac{\sum y}{N} = \frac{16.9}{5} = 3.38; b_1 = \frac{\sum y \phi_1}{\sum \phi_1^2} = \frac{13.3}{10} = 1.33$$

$$\phi_1(x) = \lambda_1 \xi = 1 \cdot (x - 2) = x - 2$$

Substituting in (\*\*), the required straight line is

$$y = 3.38 + 1.33(x - 2)$$

$$\Rightarrow y = 1.33x + 0.72$$

**Example 9.9.** Fit a second degree parabola to the following data, using the method of orthogonal polynomials.

$x$	0.5	1.0	1.5	2.0	2.5	3.0
$y$	72	110	158	214	290	380

**Solution.** Let the second degree parabola be

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

and its orthogonal polynomial transform be :

$$y = b_0 + b_1 \phi_1(x) + b_2 \phi_2(x) \quad \dots(**)$$

Here we have  $N = 6$ . Let us transform to

$$\xi = \frac{x - \frac{1}{2}(1.5 + 2.0)}{\frac{1}{2}(0.5)} = 4(x - 1.75) = 4x - 7,$$

so that  $\sum \xi = 0$ . From Fisher's tables we note the values of  $\phi_1$  and  $\phi_2$  (as given in the following table) and also

$$\sum \phi_1^2 = 70, \sum \phi_2^2 = 84; \lambda_1 = 2, \lambda_2 = 3/2$$

$x$	$\xi = 4x - 7$	$y$	$\phi_1$	$\phi_2$	$y \phi_1$	$y \phi_2$
0.5	-5	72	-5	5	-360	360
1.0	-3	110	-3	-1	-330	-110

1.5	-1	158	-1	-4	-158.	-632
2.0	1	214	1	-4	214	-856
2.5	3	290	3	-1	870	-290
3.0	5	380	5	5	1900	1900
Total		1224			2136	372

$$b_0 = \frac{\sum y}{N} = \frac{1224}{6} = 204; \quad b_1 = \frac{\sum y \phi_1}{\sum \phi_1^2} = \frac{2136}{70} = 30.51$$

$$b_2 = \frac{\sum y \phi_2}{\sum \phi_2^2} = \frac{372}{84} = 4.43; \quad \phi_1(x) = \lambda \cdot r = 2[4x - 7] = 8x - 14$$

$$\begin{aligned}\phi_2(x) &= \lambda_2 \left[ \xi^2 - \frac{N^2 - 1}{12} \right] = \frac{3}{2} \left[ (4x - 7)^2 - \frac{36 - 1}{12} \right] \\ &= \frac{3}{2} \left[ 16x^2 + 49 - 56x - \frac{35}{12} \right] = 24x^2 - 84x + 69.125\end{aligned}$$

Substituting in (\*\*), we get

$$\begin{aligned}y &= 204 + 30.51(8x - 14) + 4.43(24x^2 - 84x + 69.125) \\ &= 106.32x^2 - 128.04x + 83.08\end{aligned}$$

which is the required second degree parabola of best fit.

## Correlation and Regression

---

**10-1. Bivariate Distribution, Correlation.** So far we have confined ourselves to univariate distributions, *i.e.*, the distributions involving only one variable. We may, however, come across certain series where each term of the series may assume the values of two or more variables. For example, if we measure the heights and weights of a certain group of persons, we shall get what is known as *Bivariate distribution*—one variable relating to height and other variable relating to weight.

In a bivariate distribution we may be interested to find out if there is any correlation or covariation between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables deviate in the same direction, *i.e.*, if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be *direct* or *positive*. But if they constantly deviate in the opposite directions, *i.e.*, if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be *diverse* or *negative*. For example, the correlation between (i) the heights and weights of a group of persons, (ii) the income and expenditure is positive and the correlation between (i) price and demand of a commodity, (ii) the volume and pressure of a perfect gas, is negative. Correlation is said to be *perfect* if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

**10-2. Scatter Diagram.** It is the simplest way of the diagrammatic representation of bivariate data. Thus for the bivariate distribution  $(x_i, y_i); i = 1, 2, \dots, n$ , if the values of the variables  $X$  and  $Y$  be plotted along the  $x$ -axis and  $y$ -axis respectively in the  $xy$  plane, the diagram of dots so obtained is known as *scatter diagram*. From the scatter diagram, we can form a fairly good, though vague, idea whether the variables are correlated or not, *e.g.*, if the points are very dense, *i.e.*, very close to each other, we should expect a fairly good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected. This method, however, is not suitable if the number of observations is fairly large.

**10-3. Karl Pearson Coefficient of Correlation.** As a measure of intensity or degree of linear relationship between two variables, Karl Pearson (1867—1936), a British Biometrician, developed a formula called *Correlation Coefficient*.

Correlation coefficient between two random variables  $X$  and  $Y$ , usually denoted by  $r(X, Y)$  or simply  $r_{XY}$ , is a numerical measure of *linear relationship* between them and is defined as

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad (10-1)$$

If  $(x_i, y_i) ; i = 1, 2, \dots, n$  is the bivariate distribution, then

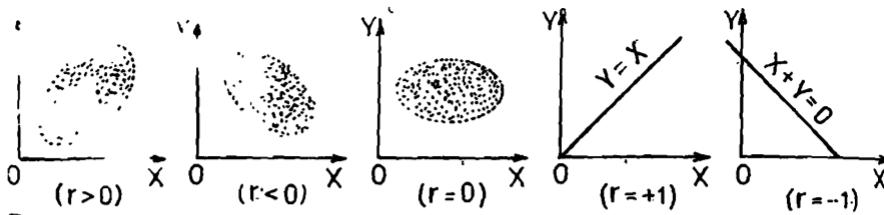
$$\left. \begin{aligned} \text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = \mu_{11} \\ \sigma_X^2 &= E(X - E(X))^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \\ \sigma_Y^2 &= E(Y - E(Y))^2 = \frac{1}{n} \sum (y_i - \bar{y})^2 \end{aligned} \right\}, \quad \dots (10.2)$$

the summation extending over  $i$  from 1 to  $n$ .

Another convenient form of the formula (10.2) for computational work is as follows :

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} (x_1 y_1 - \bar{x} \bar{y} - \bar{x} y_1 + \bar{x} \bar{y}) \\ &= \frac{1}{n} \sum x_i y_i - \bar{y} \frac{1}{n} \sum x_i - \bar{x} \frac{1}{n} \sum y_i + \bar{x} \bar{y} \\ \therefore \text{Cov}(X, Y) &= \frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}, \sigma_X^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2 \\ \text{and } \sigma_Y^2 &= \frac{1}{n} \sum y_i^2 - \bar{y}^2 \end{aligned} \quad \dots (10.2a)$$

**Remarks 1.** Following are the figures of the standard data for  $r > 0$ ,  $< 0$ ,  $= 0$ , and  $r = \pm 1$ .



2. It may be noted that  $r(X, Y)$  provides a measure of *linear relationship* between  $X$  and  $Y$ . For nonlinear relationship, however, it is not very suitable.

3. Sometimes, we write :  $\text{Cov}(X, Y) = \sigma_{XY}$

4. Karl Pearson's correlation coefficient is also called *product-moment correlation coefficient*, since

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = \mu_{11}.$$

**10.3.1. Limits for Correlation Coefficient.** We have

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\left[ \frac{1}{n} \sum (x_i - \bar{x})^2 \cdot \frac{1}{n} \sum (y_i - \bar{y})^2 \right]^{1/2}},$$

$$\therefore r^2(X, Y) = \frac{(\sum a_i b_i)^2}{(\sum a_i^2)(\sum b_i^2)}, \text{ where } \begin{cases} a_i = x_i - \bar{x} \\ b_i = y_i - \bar{y} \end{cases} \quad \dots(*)$$

We have the Schwartz inequality which states that if  $a_i, b_i; i = 1, 2, \dots, n$  are real quantities then

$$(\sum_{i=1}^n a_i b_i)^2 \leq (\sum_{i=1}^n a_i^2)(\sum_{i=1}^n b_i^2)$$

the sign of equality holding if and only if

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \dots = \frac{a_n}{b_n}$$

Using Schwartz inequality, we get from (\*)

$$r^2(X, Y) \leq 1 \text{ i.e., } |r(X, Y)| \leq 1 \Rightarrow -1 \leq r(X, Y) \leq 1 \quad \dots(10-3)$$

Hence correlation coefficient cannot exceed unity numerically. It always lies between -1 and +1. If  $r = +1$ , the correlation is perfect and positive and if  $r = -1$ , correlation is perfect and negative.

Aliter. If we write  $E(X) = \mu_X$  and  $E(Y) = \mu_Y$ , then we have

$$\begin{aligned} & E \left[ \left( \frac{X - \mu_X}{\sigma_X} \right) \pm \left( \frac{Y - \mu_Y}{\sigma_Y} \right) \right]^2 \geq 0 \\ \Rightarrow & E \left( \frac{X - \mu_X}{\sigma_X} \right)^2 + E \left( \frac{Y - \mu_Y}{\sigma_Y} \right)^2 \pm 2 \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \geq 0 \\ \Rightarrow & 1 + 1 \pm 2r(X, Y) \geq 0 \\ \Rightarrow & -1 \leq r(X, Y) \leq 1. \end{aligned}$$

**Theorem 10-1.** Correlation coefficient is independent of change of origin and scale.

**Proof.** Let  $U = \frac{X - a}{h}, V = \frac{Y - b}{k}$ , so that  $X = a + hU$  and  $Y = b + kV$ , where  $a, b, h, k$  are constants;  $h > 0, k > 0$ .

We shall prove that  $r(X, Y) = r(U, V)$

Since  $X = a + hU$  and  $Y = b + kV$ , on taking expectations, we get

$$\begin{aligned} E(X) &= a + hE(U) \quad \text{and} \quad E(Y) = b + kE(V) \\ \therefore X - E(X) &= h[U - E(U)] \quad \text{and} \quad Y - E(Y) = k[V - E(V)] \\ \Rightarrow \text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E[h(U - E(U))(k(V - E(V)))] \\ &= hk E[(U - E(U))(V - E(V))] = hk \text{ Cov}(U, V) \quad \dots(10-4) \\ \sigma_X^2 &= E[(X - E(X))^2] = E[h^2(U - E(U))^2] = h^2 \sigma_U^2 \\ \Rightarrow \sigma_X &= h \sigma_U, (h > 0) \quad \dots(10-4a) \\ \text{and} \quad \sigma_Y^2 &= E[(Y - E(Y))^2] = E[k^2(V - E(V))^2] = k^2 \sigma_V^2 \\ \Rightarrow \sigma_Y &= k \sigma_V, (k > 0) \quad \dots(10-4b) \end{aligned}$$

Substituting from (10-4), (10-4a) and (10-4b) in (10-1), we get

$$r(X, Y) = \frac{\text{Cov}(\bar{X}, Y)}{\sigma_X \sigma_Y} = \frac{hk \cdot \text{Cov}(U, V)}{hk \cdot \sigma_U \sigma_V} = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = r(U, V)$$

This theorem is of fundamental importance in the numerical computation of the correlation coefficient.

**Corollary.** If  $X$  and  $Y$  are random variables and  $a, b, c, d$  are any numbers provided only that  $a \neq 0, c \neq 0$ , then

$$r(aX + b, cY + d) = \frac{ac}{|ac|} r(X, Y)$$

**Proof.** With usual notations, we have

$$\text{Var}(aX + b) = a^2 \sigma_X^2; \quad \text{Var}(cY + d) = c^2 \sigma_Y^2;$$

$$\text{Cov}(aX + b, cY + d) = ac \sigma_{XY}$$

$$\therefore r(aX + b, cY + d) = \frac{\text{Cov}(aX + b, cY + d)}{[\text{Var}(aX + b) \text{Var}(cY + d)]^{1/2}} \\ = \frac{ac \sigma_{XY}}{|a||c|\sigma_X \sigma_Y} = \frac{ac}{|ac|} r(X, Y)$$

If  $ac > 0$ , i.e., if  $a$  and  $c$  are of same signs, then  $ac/|ac| = +1$

If  $ac < 0$ , i.e., if  $a$  and  $c$  are of opposite signs, then  $ac/|ac| = -1$ .

**Theorem 10-2.** Two independent variables are uncorrelated.

**Proof.** If  $X$  and  $Y$  are independent variables, then

$$\text{Cov}(X, Y) = 0 \quad (\text{c.f. } \S\ 6-4)$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Hence two independent variables are uncorrelated.

But the converse of the theorem is not true, i.e., two uncorrelated variables may not be independent as the following example illustrates :

$X$	-3	-2	-1	1	2	3	Total $\sum X = 0$
$Y$	9	4	1	1	4	9	$\sum Y = 28$
$XY$	-27	-8	-1	1	8	27	$\sum XY = 0$

$$\bar{X} = \frac{1}{n} \sum X = 0, \quad \text{Cov}(X, Y) = \frac{1}{n} \sum XY - \bar{X} \bar{Y} = 0$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Thus in the above example, the variables  $X$  and  $Y$  are uncorrelated. But on careful examination we find that  $X$  and  $Y$  are not independent but they are connected by the relation  $Y = X^2$ . Hence two uncorrelated variables need not necessarily be independent. A simple reasoning for this strange conclusion is that  $r(X, Y) = 0$ , merely implies the absence of any linear relationship between

the variables  $X$  and  $Y$ . There may, however, exist some other form of relationship between them, e.g., quadratic, cubic or trigonometric.

**Remarks.** 1. Following are some more examples where two variables are uncorrelated but not independent.

$$(i) X \sim N(0, 1) \text{ and } Y = X^2$$

$$\text{Since } X \sim N(0, 1), E(X) = 0 = E(X^3)$$

$$\therefore \text{Cov}(X, Y) = E(XY) - E(X)E(Y) \\ = E(X^3) - E(X)E(Y) = 0 \quad (\because Y = X^2)$$

$$\Rightarrow r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Hence  $X$  and  $Y$  are uncorrelated but not independent.

(ii) Let  $X$  be a r.v. with p.d.f.

$$f(x) = \frac{1}{2}, -1 \leq x \leq 1$$

and let  $Y = X^2$ . Here we shall get

$$E(X) = 0 \text{ and } E(XY) = E(X^3) = 0, \Rightarrow r(X, Y) = 0$$

2. However, the converse of the theorem holds in the following cases :

(a) If  $X$  and  $Y$  are jointly normally distributed with  $\rho = \rho(X, Y) = 0$ , then they are independent. If  $\rho = 0$ , then [c.f. § 10.10, Equation (10.25)]

$$f(x, y) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X - \mu_X}{\sigma_X}\right)^2\right] \times \frac{1}{\sigma_Y \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y - \mu_Y}{\sigma_Y}\right)^2\right]$$

$$\therefore f(x, y) = f_1(x)f_2(y)$$

$$\Rightarrow X \text{ and } Y \text{ are independent.}$$

(b) If each of the two variables  $X$  and  $Y$  takes two values, 0, 1 with positive probabilities, then  $r(X, Y) = 0 \Rightarrow X \text{ and } Y \text{ are independent.}$

**Proof.** Let  $X$  take the values 1 and 0 with positive probabilities  $p_1$  and  $q_1$  respectively and let  $Y$  take the values 1 and 0 with positive probabilities  $p_2$  and  $q_2$  respectively. Then

$$\begin{aligned} r(X, Y) &= 0, \Rightarrow \text{Cov}(X, Y) = 0 \\ \Rightarrow 0 &= E(XY) - E(X)E(Y) \\ &= 1 \cdot P(X = 1 \cap Y = 1) - [1 \cdot P(X = 1) \times 1 \cdot P(Y = 1)] \\ &= P(X = 1 \cap Y = 1) - p_1 p_2 \\ \Rightarrow P(X = 1 \cap Y = 1) &= p_1 p_2 = P(X = 1) \cdot P(Y = 1) \\ \Rightarrow X \text{ and } Y \text{ are independent.} \end{aligned}$$

**10.3.2. Assumptions Underlying Karl Pearson's Correlation Coefficient.** Pearsonian correlation coefficient  $r$  is based on the following assumptions :

(i) The variables  $X$  and  $Y$  under study are linearly related. In other words, the scatter diagram of the data will give a straight line curve.

(ii) Each of the variables (series) is being affected by a large number of independent contributory causes of such a nature as to produce normal distribution. For example, the variables (series) relating to ages, heights, weight, supply, price, etc., conform to this assumption. In the words of Karl Pearson :

"The sizes of the complex of organs (something measurable) are determined by a great variety of independent contributory causes, for example, climate, nourishment, physical training and innumerable other causes which cannot be individually observed or their effects measured." Karl Pearson further observes, "The variations in intensity of the contributory causes are small as compared with their absolute intensity and these variations follow the normal law of distribution."

(iii) The forces so operating on each of the variable series are not independent of each other but are related in a causal fashion. In other word, cause and effect relationship exists between different forces operating on the items of the two variable series. These forces must be common to both the series. If the operating forces are entirely independent of each other and not related in any fashion, then there cannot be any correlation between the variables under study.

For example, the correlation coefficient between,

- (a) the series of heights and incomes of individuals over a period of time,
- (b) the series of marriage rate and the rate of agricultural production in a country over a period of time,

(c) the series relating to the size of the shoe and intelligence of a group of individuals,

should be zero, since the forces affecting the two variable series in each of the above cases are entirely independent of each other.

However, if in any of the above cases the value of  $r$  for a given set of data is not zero, then such correlation is termed as *chance correlation* or *spurious* or *nonsense correlation*.

**Example 10.1.** Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y) :

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

**Solution.**

#### CALCULATIONS FOR CORRELATION COEFFICIENT

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

$$\bar{X} = \frac{1}{n} \sum X = \frac{544}{8} = 68, \bar{Y} = \frac{1}{n} \sum Y = \frac{1}{8} \times 552 = 69$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n} \sum XY - \bar{X} \bar{Y}}{\sqrt{\left(\frac{1}{n} \sum X^2 - \bar{X}^2\right) \left(\frac{1}{n} \sum Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{\frac{1}{8} \times 37560 - 68 \times 69}{\sqrt{\left[\frac{37028}{8} - (68)^2\right] \left[\frac{38132}{8} - (69)^2\right]}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628.5 - 4624)(4766.5 - 4761)}} = \frac{3}{\sqrt{4.5 \times 5.5}} = 0.603$$

Aliter.

(SHORT-CUT METHOD)

X	Y	$U = X - 68$	$V = Y - 69$	$U^2$	$V^2$	$UV$
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total		0	0	36	44	24

$$\bar{U} = \frac{1}{n} \sum U = 0, \bar{V} = \frac{1}{n} \sum V = 0$$

$$\text{Cov}(U, V) = \frac{1}{n} \sum UV - \bar{U} \bar{V} = \frac{1}{8} \times 24 = 3$$

$$\sigma_U^2 = \frac{1}{n} \sum U^2 - (\bar{U})^2 = \frac{1}{8} \times 36 = 4.5$$

$$\sigma_V^2 = \frac{1}{n} \sum V^2 - (\bar{V})^2 = \frac{1}{8} \times 44 = 5.5$$

$$\therefore r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{3}{\sqrt{4.5 \times 5.5}} = 0.603 = r(X, Y)$$

**Remark.** The reader is advised to calculate the correlation coefficient by arbitrary origin method rather than by the direct method; since the latter leads to much simpler arithmetical calculations.

**Example 10-2.** A computer while calculating correlation coefficient between two variables  $X$  and  $Y$  from 25 pairs of observations obtained the following results :

$$n = 25, \sum X = 125, \sum X^2 = 650, \sum Y = 100, \sum Y^2 = 460, \sum XY = 508$$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

X	Y
8	12
6	8

Obtain the correct value of correlation coefficient.

[Calcutta Univ. B.Sc. (Maths. Hons.), 1988, 1991]

**Solution.**

$$\text{Corrected } \sum X = 125 - 6 - 8 + 8 + 6 = 125$$

$$\text{Corrected } \sum Y = 100 - 14 - 6 + 12 + 8 = 100$$

$$\text{Corrected } \sum X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$$

$$\text{Corrected } \sum Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$$

$$\text{Corrected } \sum XY = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$$

$$\bar{X} = \frac{1}{n} \sum X = \frac{1}{25} \times 125 = 5, \quad \bar{Y} = \frac{1}{n} \sum Y = \frac{1}{25} \times 100 = 4$$

$$\text{Cov}(X, Y) = \frac{1}{n} \sum XY - \bar{X}\bar{Y} = \frac{1}{25} \times 520 - 5 \times 4 = \frac{4}{5}$$

$$\sigma_X^2 = \frac{1}{n} \sum X^2 - \bar{X}^2 = \frac{1}{25} \times 650 - (5)^2 = 1$$

$$\sigma_Y^2 = \frac{1}{n} \sum Y^2 - \bar{Y}^2 = \frac{1}{25} \times 436 - 16 = \frac{36}{25}$$

$$\therefore \text{Corrected } r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{4}{5}}{1 \times \frac{6}{5}} = \frac{2}{3} = 0.67$$

**Example 10-3.** Show that if  $X'$ ,  $Y'$  are the deviations of the random variables  $X$  and  $Y$  from their respective means then

$$(i) \quad r = 1 - \frac{1}{2N} \sum_i \left( \frac{X'_i}{\sigma_X} - \frac{Y'_i}{\sigma_Y} \right)^2$$

$$(ii) \quad r = -1 + \frac{1}{2N} \sum_i \left( \frac{X'_i}{\sigma_X} + \frac{Y'_i}{\sigma_Y} \right)^2$$

Deduce that  $-1 \leq r \leq +1$ .

[Delhi Univ. B.Sc. Oct. 1992; Madras Univ. B.Sc., Nov. 1991]

**Solution.** (i) Here  $X'_i = (x_i - \bar{X})$  and  $Y'_i = (Y_i - \bar{Y})$

$$\text{R.H.S.} = 1 - \frac{1}{2N} \sum_i \left( \frac{X'_i}{\sigma_X} - \frac{Y'_i}{\sigma_Y} \right)^2$$

$$\begin{aligned}
 &= 1 - \frac{1}{2N} \sum_i \left[ \frac{X_i^2}{\sigma_X^2} + \frac{Y_i^2}{\sigma_Y^2} - \frac{2X_i Y_i}{\sigma_X \sigma_Y} \right] \\
 &= 1 - \frac{1}{2N} \left[ \frac{1}{\sigma_X^2} \sum_i X_i^2 + \frac{1}{\sigma_Y^2} \sum_i Y_i^2 - \frac{2}{\sigma_X \sigma_Y} \sum_i X_i Y_i \right] \\
 &= 1 - \frac{1}{2N} \left[ \frac{1}{\sigma_X^2} \sum_i (X_i - \bar{X})^2 + \frac{1}{\sigma_Y^2} \sum_i (Y_i - \bar{Y})^2 - \frac{2}{\sigma_X \sigma_Y} \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) \right] \\
 &= 1 - \frac{1}{2} \left[ \frac{1}{\sigma_X^2} \cdot \sigma_X^2 + \frac{1}{\sigma_Y^2} \cdot \sigma_Y^2 - \frac{2}{\sigma_X \sigma_Y} \cdot r \sigma_X \sigma_Y \right] \\
 &= 1 - \frac{1}{2} [1 + 1 - 2r] = r
 \end{aligned}$$

(ii) Proceeding similarly, we will get

$$\text{R.H.S.} = -1 + \frac{1}{2}(1 + 1 + 2r) = r$$

**Deduction.** Since  $\left(\frac{X_i}{\sigma_X} \pm \frac{Y_i}{\sigma_Y}\right)^2$ , being the square of a real quantity is always non-negative,  $\sum_i \left(\frac{X_i}{\sigma_X} \mp \frac{Y_i}{\sigma_Y}\right)^2$  is also non-negative. From part (i), we get

$$r = 1 - (\text{some non-negative quantity}) \Rightarrow r \leq 1 \quad \dots(*)$$

Also from part (ii), we get

$$r = -1 + (\text{some non-negative quantity}) \Rightarrow -1 \leq r \quad \dots(**)$$

The sign of equality in (\*) and (\*\*) holds if and only if

$$\left. \begin{array}{l} \frac{X_i}{\sigma_X} - \frac{Y_i}{\sigma_Y} = 0 \\ \frac{X_i}{\sigma_X} + \frac{Y_i}{\sigma_Y} = 0 \end{array} \right\} \forall i = 1, 2, \dots, n$$

and

respectively.

From (\*) and (\*\*), we get

$$-1 \leq r \leq 1$$

**Example 10-4.** The variables  $X$  and  $Y$  are connected by the equation  $aX + bY + c = 0$ . Show that the correlation between them is  $-1$  if the signs of  $a$  and  $b$  are alike and  $+1$  if they are different.

[Nagpur Univ. B.Sc. 1992; Delhi Univ. B.Sc. (Stat. Hons.) 1992]

**Solution.**  $aX + bY + c = 0 \Rightarrow aE(X) + bE(Y) + c = 0$

$$\therefore a(X - E(X)) + b(Y - E(Y)) = 0$$

$$\Rightarrow (X - E(X)) = -\frac{b}{a} (Y - E(Y))$$

$$\therefore \text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

$$\begin{aligned}
 &= -\frac{b}{a} E[(Y - E(Y))^2] = -\frac{b}{a} \cdot \sigma_Y^2 \\
 E(X - E(X))^2 &= \frac{b^2}{a^2} E[(Y - E(Y))^2] = \frac{b^2}{a^2} \cdot \sigma_Y^2 \\
 \therefore r &= \frac{-\frac{b}{a} \cdot \sigma_Y^2}{\sqrt{\sigma_Y^2} \sqrt{\frac{b^2}{a^2} \cdot \sigma_Y^2}} = \frac{-\frac{b}{a} \sigma_Y^2}{\left| \frac{b}{a} \right| \sigma_Y^2} \\
 &= \begin{cases} +1, & \text{if } b \text{ and } a \text{ are of opposite signs.} \\ -1, & \text{if } b \text{ and } a \text{ are of same sign.} \end{cases}
 \end{aligned}$$

**Example 10-5.** (a) If  $Z = aX + bY$  and  $r$  is the correlation coefficient between  $X$  and  $Y$ , show that

$$\sigma_Z^2 = a^2 \sigma_X^2 + b^2 \sigma_Y^2 + 2ab r \sigma_X \sigma_Y$$

(b) Show that the correlation coefficient  $r$  between two random variables  $X$  and  $Y$  is given by

$$r = (\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2) / 2\sigma_X \sigma_Y$$

where  $\sigma_X$ ,  $\sigma_Y$  and  $\sigma_{X-Y}$  are the standard deviations of  $X$ ,  $Y$  and  $X - Y$  respectively.

[Calcutta Univ. B.Sc., 1992; M.S. Baroda Univ. B.Sc. 1992]

**Solution.** Taking expectation of both sides of  $Z = aX + bY$ , we get

$$E(Z) = aE(X) + bE(Y)$$

$$\therefore Z - E(Z) = a(X - E(X)) + b(Y - E(Y))$$

Squaring and taking expectation of both sides, we get

$$\begin{aligned}
 \sigma_Z^2 &= a^2 \sigma_X^2 + b^2 \sigma_Y^2 + 2ab \operatorname{Cov}(X, Y) \\
 &= a^2 \sigma_X^2 + b^2 \sigma_Y^2 + 2ab r \sigma_X \sigma_Y
 \end{aligned}$$

(b) Taking  $a = 1$ ,  $b = -1$  in the above case, we have

$$Z = X - Y \text{ and } \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r \sigma_X \sigma_Y$$

$$\therefore r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X \sigma_Y}$$

**Remark.** In the above example, we have obtained

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab \operatorname{Cov}(X, Y)$$

Similarly, we could obtain the result

$$V(aX - bY) = a^2 V(X) + b^2 V(Y) - 2ab \operatorname{Cov}(X, Y)$$

The above results are useful in solving theoretical problems.

**Example 10-6.**  $X$  and  $Y$  are two random variables with variances  $\sigma_X^2$  and  $\sigma_Y^2$  respectively and  $r$  is the coefficient of correlation between them. If

$U = X + kY$  and  $V = X + \frac{\sigma_X}{\sigma_Y} Y$ , find the value of  $k$  so that  $U$  and  $V$  are uncorrelated.

[Delhi Univ. B.Sc. 1992; Andhra Univ. B.Sc. 1993]

**Solution.** Taking expectations of  $U = X + kY$  and  $V = X + \frac{\sigma_X}{\sigma_Y} Y$ , we get

$$E(U) = E(X) + kE(Y) \text{ and } E(V) = E(X) + \frac{\sigma_X}{\sigma_Y} E(Y)$$

$$U - E(U) = (X - E(X)) + k(Y - E(Y)) \text{ and}$$

$$V - E(V) = (X - E(X)) + \frac{\sigma_X}{\sigma_Y} (Y - E(Y))$$

$$\text{Cov}(U, V) = E[(U - E(U))(V - E(V))]$$

$$\begin{aligned} &= E[(X - E(X)) + k(Y - E(Y))] \times [(X - E(X)) + \frac{\sigma_X}{\sigma_Y} (Y - E(Y))] \\ &= \sigma_X^2 + \frac{\sigma_X}{\sigma_Y} \text{Cov}(X, Y) + k \text{Cov}(X, Y) + k \frac{\sigma_X}{\sigma_Y} \cdot \sigma_Y^2 \\ &= \left[ \sigma_X^2 + k\sigma_X\sigma_Y \right] + \left[ \frac{\sigma_X}{\sigma_Y} + k \right] \text{Cov}(X, Y) \\ &= \sigma_X(\sigma_X + k\sigma_Y) + \left[ \frac{\sigma_X + k\sigma_Y}{\sigma_Y} \right] \text{Cov}(X, Y) \\ &= (\sigma_X + k\sigma_Y) \left[ \sigma_X + \frac{\text{Cov}(X, Y)}{\sigma_Y} \right] = (\sigma_X + k\sigma_Y)(1 + r)\sigma_X \end{aligned}$$

$U$  and  $V$  will be uncorrelated if

$$r(U, V) = 0 \Rightarrow \text{Cov}(U, V) = 0$$

$$\text{i.e., if } (\sigma_X + k\sigma_Y)(1 + r)\sigma_X = 0$$

$$\Rightarrow \sigma_X + k\sigma_Y = 0 \quad (\because \sigma_X \neq 0, r \neq -1)$$

$$\Rightarrow k = -\frac{\sigma_X}{\sigma_Y}$$

**Example 10.7.** The random variables  $X$  and  $Y$  are jointly normally distributed and  $U$  and  $V$  are defined by

$$U = X \cos \alpha + Y \sin \alpha,$$

$$V = Y \cos \alpha - X \sin \alpha$$

Show that  $U$  and  $V$  will be uncorrelated if

$$\tan 2\alpha = \frac{2r\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2},$$

where  $r = \text{corr.}(X, Y)$ ,  $\sigma_X^2 = \text{Var}(X)$  and  $\sigma_Y^2 = \text{Var}(Y)$ . Are  $U$  and  $V$  then independent?

[Delhi Univ. B.Sc. (Stat. Hons.) 1989; (Maths. Hons.), 1990]

**Solution.** We have

$$\text{Cov}(U, V) = E[(U - E(U))(V - E(V))]$$

$$\begin{aligned} &= E[(X - E(X)) \cos \alpha + (Y - E(Y)) \sin \alpha] \\ &\quad \times [(Y - E(Y)) \cos \alpha - (X - E(X)) \sin \alpha] \end{aligned}$$

$$\begin{aligned}
 &= \cos^2 \alpha \operatorname{Cov}(X, Y) - \sin \alpha \cos \alpha \cdot \sigma_X^2 \\
 &\quad + \sin \alpha \cos \alpha \cdot \sigma_Y^2 - \sin^2 \alpha (\operatorname{Cov}(X, Y)) \\
 &= (\cos^2 \alpha - \sin^2 \alpha) \operatorname{Cov}(X, Y) - \sin \alpha \cos \alpha (\sigma_X^2 - \sigma_Y^2) \\
 &= \cos 2\alpha \operatorname{Cov}(X, Y) - \sin \alpha \cos \alpha (\sigma_X^2 - \sigma_Y^2)
 \end{aligned}$$

$U$  and  $V$  will be uncorrelated if and only if

$$r(U, V) = 0, \text{ i.e., iff } \operatorname{Cov}(U, V) = 0$$

$$\text{i.e., if } \cos 2\alpha \operatorname{Cov}(X, Y) - \sin \alpha \cos \alpha (\sigma_X^2 - \sigma_Y^2) = 0$$

$$\text{or if } \cos 2\alpha r \sigma_X \sigma_Y = \frac{\sin 2\alpha}{2} (\sigma_X^2 - \sigma_Y^2)$$

$$\text{or if } \tan 2\alpha = \frac{2r \sigma_X \sigma_Y}{\sigma_X^2 - \sigma_Y^2}$$

However,  $r(U, V) = 0$  does not imply that the variables  $U$  and  $V$  are independent. [For detailed discussion, see Theorem 10-2, page 10-4.]

**Example 10-8.** If  $X, Y$  are standardized random variables, and

$$r(aX + bY, bX + aY) = \frac{1 + 2ab}{a^2 + b^2} \quad \dots(*)$$

find  $r(X, Y)$ , the coefficient of correlation between  $X$  and  $Y$ .

[Sardar Patel Univ. B.Sc., 1993; Delhi Univ. B.Sc. (Stat. Hons.), 1989]

**Solution.** Since  $X$  and  $Y$  are standardised random variables, we have

$$\begin{aligned}
 \text{and } E(X) &= E(Y) = 0 \\
 \text{and } \operatorname{Var}(X) &= \operatorname{Var}(Y) = 1 \Rightarrow \bar{E}(X^2) = E(Y^2) = 1 \\
 \text{and } \operatorname{Cov}(X, Y) &= E(XY) \Rightarrow E(XY) = r(X, Y) \cdot \sigma_X \sigma_Y \\
 &\qquad\qquad\qquad = r(X, Y)
 \end{aligned} \quad \dots(**)$$

Also we have

$$\begin{aligned}
 &r(aX + bY, bX + aY) \\
 &= \frac{E[(aX + bY)(bX + aY)] - E(aX + bY) E(bX + aY)}{[\operatorname{Var}(aX + bY) \cdot \operatorname{Var}(bX + aY)]^{1/2}} \\
 &= \frac{E[abX^2 + a^2 XY + b^2 YX + abY^2] - 0}{\{[a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X, Y)]} \\
 &\quad \times [b^2 \operatorname{Var}(X) + a^2 \operatorname{Var}(Y) + 2ba \operatorname{Cov}(X, Y)]\}^{1/2} \\
 &= \frac{ab \cdot 1 + a^2 r(X, Y) + b^2 r(X, Y) + ab \cdot 1}{\{[a^2 + b^2 + 2ab r(X, Y)][b^2 + a^2 + 2ba r(X, Y)]\}^{1/2}} \\
 &\qquad\qquad\qquad \text{[Using (**)]} \\
 &= \frac{2ab + (a^2 + b^2) \cdot r(X, Y)}{a^2 + b^2 + 2ab \cdot r(X, Y)}
 \end{aligned}$$

From (\*) and (\*\*), we get

$$\frac{1 + 2ab}{a^2 + b^2} = \frac{(a^2 + b^2) \cdot r(X, Y) + 2ab}{a^2 + b^2 + 2ab \cdot r(X, Y)}$$

Cross multiplying, we get

$$\begin{aligned}
 & (a^2 + b^2)(1 + 2ab) + 2ab \cdot r(X, Y)(1 + 2ab) = (a^2 + b^2)^2 \cdot r(X, Y) + 2ab(a^2 + b^2) \\
 \Rightarrow & (a^4 + b^4 + 2a^2b^2 - 2ab - 4a^2b^2) \cdot r(X, Y) = (a^2 + b^2)^2 \\
 \Rightarrow & [(a^2 - b^2)^2 - 2ab]r(X, Y) = a^2 + b^2 \\
 \Rightarrow & r(X, Y) = \frac{a^2 + b^2}{(a^2 - b^2)^2 - 2ab}
 \end{aligned}$$

**Example 10.9.** If  $X$  and  $Y$  are uncorrelated random variables with means zero and variances  $\sigma_1^2$  and  $\sigma_2^2$  respectively, show that

$$U = X \cos \alpha + Y \sin \alpha, V = X \sin \alpha - Y \cos \alpha$$

have a correlation coefficient  $\rho$  given by

$$\rho = \frac{\sigma_1^2 - \sigma_2^2}{[(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_1^2\sigma_2^2 \operatorname{cosec}^2 2\alpha]^{1/2}}$$

**Solution.** We are given that

$$r(X, Y) = 0 \Rightarrow \operatorname{Cov}(X, Y) = 0, \sigma_X^2 = \sigma_1^2 \text{ and } \sigma_Y^2 = \sigma_2^2 \quad \dots(1)$$

We have

$$\begin{aligned}
 \sigma_U^2 &= V(X \cos \alpha + Y \sin \alpha) \\
 &= \cos^2 \alpha V(X) + \sin^2 \alpha V(Y) + 2 \sin \alpha \cos \alpha \operatorname{Cov}(X, Y) \\
 &= \cos^2 \alpha \sigma_1^2 + \sin^2 \alpha \sigma_2^2 \quad [\text{Using (1)}]
 \end{aligned}$$

Similarly,

$$\sigma_V^2 = V(X \sin \alpha - Y \cos \alpha) = \sin^2 \alpha \cdot \sigma_1^2 + \cos^2 \alpha \cdot \sigma_2^2$$

$$\begin{aligned}
 \operatorname{Cov}(U, V) &= E[(U - E(U))(V - E(V))] \\
 &= E \left[ ((X - E(X)) \cos \alpha + (Y - E(Y)) \sin \alpha) \right. \\
 &\quad \times \left. ((X - E(X)) \sin \alpha - (Y - E(Y)) \cos \alpha) \right] \\
 &= \sin \alpha \cos \alpha V(X) - \cos^2 \alpha \operatorname{Cov}(X, Y) \\
 &\quad + \sin^2 \alpha \operatorname{Cov}(X, Y) - \sin \alpha \cos \alpha V(Y) \\
 &= (\sigma_1^2 - \sigma_2^2) \sin \alpha \cos \alpha \quad [\text{Using (1)}]
 \end{aligned}$$

Now

$$\rho^2 = \frac{[\operatorname{Cov}(U, V)]^2}{\sigma_U^2 \sigma_V^2}$$

$$\begin{aligned}
 \text{where } \sigma_U^2 \sigma_V^2 &= (\cos^2 \alpha \sigma_1^2 + \sin^2 \alpha \sigma_2^2)(\sin^2 \alpha \sigma_1^2 + \cos^2 \alpha \sigma_2^2) \\
 &= \sin^2 \alpha \cos^2 \alpha (\sigma_1^4 + \sigma_2^4) + \sigma_1^2 \sigma_2^2 (\cos^4 \alpha + \sin^4 \alpha) \\
 &= \sin^2 \alpha \cos^2 \alpha (\sigma_1^4 + \sigma_2^4) + \sigma_1^2 \sigma_2^2 [(\sin^2 \alpha + \cos^2 \alpha)^2 - 2 \sin^2 \alpha \cos^2 \alpha] \\
 &= \sin^2 \alpha \cos^2 \alpha (\sigma_1^4 + \sigma_2^4 - 2\sigma_1^2 \sigma_2^2) + \sigma_1^2 \sigma_2^2 \\
 &= \sin^2 \alpha \cos^2 \alpha (\sigma_1^2 - \sigma_2^2)^2 + \sigma_1^2 \sigma_2^2 \\
 \therefore \rho^2 &= \frac{(\sigma_1^2 - \sigma_2^2)^2 \cdot \sin^2 \alpha \cos^2 \alpha}{\sigma_1^2 \sigma_2^2 + \sin^2 \alpha \cos^2 \alpha (\sigma_1^2 - \sigma_2^2)^2} \\
 &= \frac{\frac{1}{4}(\sigma_1^2 - \sigma_2^2)^2 \sin^2 2\alpha}{\sigma_1^2 \sigma_2^2 + \sin^2 2\alpha \cdot \frac{1}{4}(\sigma_1^2 - \sigma_2^2)^2}
 \end{aligned}$$

$$= \frac{(\sigma_1^2 - \sigma_2^2)^2}{4\sigma_1^2\sigma_2^2 \cosec^2 2\alpha + (\sigma_1^2 - \sigma_2^2)^2}$$

$$\Rightarrow \rho = \frac{\sigma_1^2 - \sigma_2^2}{[(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_1^2\sigma_2^2 \cosec^2 2\alpha]^{1/2}}$$

**Example 10.10.** If  $U = aX + bY$  and  $V = cX + dY$ , where  $X$  and  $Y$  are measured from their respective means and if  $r$  is the correlation coefficient between  $X$  and  $Y$ , and if  $U$  and  $V$  are uncorrelated, show that

$$\sigma_U\sigma_V = (ad - bc)\sigma_X\sigma_Y(1 - r^2)^{1/2}$$

[Poona Univ. B.Sc., 1990; Delhi Univ. B.Sc. (Stat. Hons.), 1986]

**Solution.** We have

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \Rightarrow 1 - r^2 = 1 - \frac{[\text{Cov}(X, Y)]^2}{\sigma_X^2 \sigma_Y^2}$$

$$\Rightarrow (1 - r^2) \sigma_X^2 \sigma_Y^2 = \sigma_X^2 \sigma_Y^2 - [\text{Cov}(X, Y)]^2 \quad \dots(*)$$

[This step is suggested by the answer]

$$U = aX + bY, V = cX + dY$$

Since  $X, Y$  are measured from their means,

$$E(X) = 0 = E(Y) \Rightarrow E(U) = 0 = E(V) \quad \left. \begin{array}{l} \\ \end{array} \right\} \quad \dots(**)$$

and  $\sigma_U^2 = E(U^2); \sigma_V^2 = E(V^2)$

$$\text{Also } aX + bY - U = 0 \text{ and } cX + dY - V = 0$$

$$\Rightarrow \frac{X}{-bV + dU} = \frac{Y}{-cU + aV} = \frac{1}{ad - bc}$$

$$\Rightarrow \left. \begin{array}{l} X = \frac{1}{ad - bc} (dU - bV) \\ Y = \frac{1}{ad - bc} (-cU + aV) \end{array} \right\} \quad \dots(***)$$

$$\therefore \text{Var}(X) = \frac{1}{(ad - bc)^2} [d^2 \sigma_U^2 + b^2 \sigma_V^2 - 2bd \text{Cov}(U, V)]$$

$$= \frac{1}{(ad - bc)^2} [d^2 \sigma_U^2 + b^2 \sigma_V^2]$$

[Since  $U, V$  are uncorrelated  $\Leftrightarrow \text{Cov}(U, V) = 0$ ]

Similarly, we have

$$\text{Var}(Y) = \frac{1}{(ad - bc)^2} (c^2 \sigma_U^2 + a^2 \sigma_V^2)$$

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(XY) \quad \left[ \because E(X) = 0 = E(Y) \right]$$

$$= \frac{1}{(ad - bc)^2} E[(dU - bV)(-cU + aV)] \quad \text{[From (***)]}$$

$$\begin{aligned}
 &= \frac{1}{(ad - bc)^2} [-cd\sigma_U^2 - ab\sigma_V^2] \\
 &\quad [\text{Using } (***) \text{ and } \text{Cov}(U, V) = 0, \text{ given}] \\
 &= \frac{-1}{(ad - bc)^2} [cd\sigma_U^2 + ab\sigma_V^2]
 \end{aligned}$$

Substituting in (\*), we get

$$\begin{aligned}
 (1 - r^2)\sigma_X^2\sigma_Y^2 &= \frac{1}{(ad - bc)^4} \times [(d^2\sigma_U^2 + b^2\sigma_V^2)(c^2\sigma_U^2 + a^2\sigma_V^2) \\
 &\quad - (cd\sigma_U^2 + ab\sigma_V^2)^2] \\
 &= \frac{1}{(ad - bc)^4} \\
 &\quad \times [c^2d^2\sigma_U^4 + a^2b^2\sigma_V^4 + (a^2d^2 + b^2c^2)\sigma_U^2\sigma_V^2 \\
 &\quad - c^2d^2\sigma_U^4 - a^2b^2\sigma_V^4 - 2abcd\sigma_U^2\sigma_V^2] \\
 &= \frac{1}{(ad - bc)^4} [a^2d^2 + b^2c^2 - 2abcd]\sigma_U^2\sigma_V^2 \\
 &= \frac{1}{(ad - bc)^4} (ad - bc)^2\sigma_U^2\sigma_V^2 \\
 &= \frac{\sigma_U^2\sigma_V^2}{(ad - bc)^2}
 \end{aligned}$$

Cross multiplying and taking square root, we get the required result.

**Example 10.11. (a) Establish the formula :**

$$nr\sigma_X\sigma_Y = n_1r_1\sigma_{X_1}\sigma_{Y_1} + n_2r_2\sigma_{X_2}\sigma_{Y_2} + n_1dx_1dy_1 + n_2dx_2dy_2 \quad \dots(10.5)$$

where  $n_1$ ,  $n_2$  and  $n$  are respectively the sizes of the first, second and combined sample :  $(\bar{x}_1, \bar{y}_1)$ ,  $(\bar{x}_2, \bar{y}_2)$ ,  $(\bar{x}, \bar{y})$ , their means  $r_1$ ,  $r_2$  and  $r$  their coefficients of correlation;  $(\sigma_{X_1}, \sigma_{Y_1})$ ,  $(\sigma_{X_2}, \sigma_{Y_2})$ ,  $(\sigma_X, \sigma_Y)$  their standard deviations, and

$$dx_1 = \bar{x}_1 - \bar{x} \quad , \quad dy_1 = \bar{y}_1 - \bar{y}$$

$$dx_2 = \bar{x}_2 - \bar{x} \quad , \quad dy_2 = \bar{y}_2 - \bar{y}$$

**(b) Find the correlation co-efficient of combined sample given that**

	Sample I	Sample II
Sample size	100	150
Sample mean ( $\bar{x}$ )	80	72
Sample mean ( $\bar{y}$ )	100	118
Sample variance ( $\sigma_X^2$ )	10	12
Sample variance ( $\sigma_Y^2$ )	15	18
Correlation coefficient	0.6	0.4

**Solution.** (a) Let  $(x_{1i}, y_{1i})$ ;  $i = 1, 2, \dots, n_1$  and  $(x_{2j}, y_{2j})$ ;  $j = 1, 2, \dots, n_2$ , be the two samples of sizes  $n_1$  and  $n_2$  respectively from the bivariate population. Then with the given notations, we have

$$\begin{aligned}\bar{x} &= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}, \quad \bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2} \\ n\sigma_x^2 &= n_1 (\sigma_{x_1}^2 + dx_1^2) + n_2 (\sigma_{x_2}^2 + dx_2^2) \\ n\sigma_y^2 &= n_1 (\sigma_{y_1}^2 + dy_1^2) + n_2 (\sigma_{y_2}^2 + dy_2^2)\end{aligned}\quad \dots(1)$$

$$r_1 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1)}{n_1 \sigma_{x_1} \sigma_{y_1}}, \quad r_2 = \frac{\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(y_{2j} - \bar{y}_2)}{n_2 \sigma_{x_2} \sigma_{y_2}} \quad \dots(2)$$

For the pooled sample, we have

$$r = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x})(y_{1i} - \bar{y}) + \sum_{j=1}^{n_2} (x_{2j} - \bar{x})(y_{2j} - \bar{y})}{n \sigma_x \sigma_y} \quad \dots(3)$$

Now

$$\begin{aligned}\sum_{i=1}^{n_1} (x_{1i} - \bar{x})(y_{1i} - \bar{y}) &= \sum_{i=1}^{n_1} \left\{ \{(x_{1i} - \bar{x}_1) + (\bar{x}_1 - \bar{x})\} \{(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y})\} \right\} \\ &= \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y}) \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1) \\ &\quad + (\bar{x}_1 - \bar{x}) \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1) + n_1 (\bar{x}_1 - \bar{x})(\bar{y}_1 - \bar{y})\end{aligned}$$

$$\text{But } \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1) = 0 \text{ and } \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1) = 0,$$

being the algebraic sum of the deviations from the mean.

$$\therefore \sum_{i=1}^{n_1} (x_{1i} - \bar{x})(y_{1i} - \bar{y}) = n_1 r_1 \sigma_{x_1} \sigma_{y_1} + n_1 dx_1 dy_1 \quad [\text{Using (2)}]$$

Similarly, we will get

$$\sum_{j=1}^{n_2} (x_{2j} - \bar{x})(y_{2j} - \bar{y}) = n_2 r_2 \sigma_{x_2} \sigma_{y_2} + n_2 dx_2 dy_2$$

Substituting in (3), we get the required formula.

(b) Here we are given :

$$n_1 = 100, \bar{x}_1 = 80, \bar{y}_1 = 100, \sigma_{x_1}^2 = 10, \sigma_{y_1}^2 = 15, r_1 = 0.6$$

$$n_2 = 150, \bar{x}_2 = 72, \bar{y}_2 = 118, \sigma_{x_2}^2 = 12, \sigma_{y_2}^2 = 18, r_2 = 0.4$$

$$\therefore \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{100 \times 80 + 150 \times 72}{100 + 150} = 75.2$$

$$\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2} = \frac{100 \times 100 + 150 \times 118}{100 + 150} = 110.8$$

$$dx_1 = \bar{x}_1 - \bar{x} = 4.8, \quad dy_1 = \bar{y}_1 - \bar{y} = 10.8$$

$$dx_2 = \bar{x}_2 - \bar{x} = 3.2, \quad dy_2 = \bar{y}_2 - \bar{y} = 7.2$$

$$n\sigma_x^2 = n_1 (\sigma_{x_1}^2 + dx_1^2) + n_2 (\sigma_{x_2}^2 + dx_2^2) = 6640$$

$$n\sigma_y^2 = n_1 (\sigma_{y_1}^2 + dy_1^2) + n_2 (\sigma_{y_2}^2 + dy_2^2) = 23640$$

Substituting these values in the formula and simplifying, we get

$$r = \frac{n_1 r_1 \sigma_{x_1} \sigma_{y_1} + n_2 r_2 \sigma_{x_2} \sigma_{y_2} + n_1 dx_1 dy_1 + n_2 dx_2 dy_2}{n \sigma_x \sigma_y} = 0.8186$$

**Example 10.12.** The independent variables  $X$  and  $Y$  are defined by :

$$\begin{aligned} f(x) &= 4ax, \quad 0 \leq x \leq r \\ &= 0, \quad \text{otherwise} \end{aligned} \quad \quad \quad \begin{aligned} f(y) &= 4by, \quad 0 \leq y \leq s \\ &= 0, \quad \text{otherwise} \end{aligned}$$

Show that :

$$\text{Cov}(U, V) = \frac{b-a}{b+a},$$

$$\text{where } U = X + Y \quad \text{and} \quad V = X - Y$$

[I.I.T. (B. Tech.), Nov. 1992]

**Solution.** Since the total area under probability curve is unity (one), we have :

$$\int_0^r f(x) dx = 4a \int_0^r x dx = 1 \Rightarrow 2ar^2 = 1 \Rightarrow a = \frac{1}{2r^2} \quad \dots(i)$$

$$\int_0^s f(y) dy = 4b \int_0^s y dy = 1 \Rightarrow 2bs^2 = 1 \Rightarrow b = \frac{1}{2s^2} \quad \dots(ii)$$

$$\therefore f(x) = 4ax = \frac{2x}{r^2}, \quad 0 \leq x \leq r; \quad \text{and} \quad f(y) = 4by = \frac{2y}{s^2}, \quad 0 \leq y \leq s \quad \dots(iii)$$

Since  $X$  and  $Y$  are independent variates,

$$\text{Cov}(X, Y) = 0 \Rightarrow \text{Cov}(X, Y) = 0 \quad \dots(iv)$$

$$\text{Cov}(U, V) = \text{Cov}(X + Y, X - Y)$$

$$= \text{Cov}(X, X) - \text{Cov}(X, Y) + \text{Cov}(Y, X) - \text{Cov}(Y, Y)$$

$$= \sigma_x^2 - \sigma_y^2 \quad [\text{Using (iv)}]$$

$$\text{Var}(U) = \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$$

$$= \sigma_x^2 + \sigma_y^2 \quad [\text{Using (iv)}]$$

$$\text{Var}(V) = \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y)$$

$$= \sigma_x^2 + \sigma_y^2 \quad [\text{Using (iv)}]$$

$$\therefore r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2} \quad \dots (v)$$

We have :

$$E(X) = \int_0^r x f(x) dx = \frac{2}{r^2} \int_0^r x^2 dx = \frac{2r}{3} \quad [\text{From (iii)}]$$

$$E(X^2) = \int_0^r x^2 f(x) dx = \frac{2}{r^2} \int_0^r x^3 dx = \frac{r^2}{2}$$

$$\therefore \text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{r^2}{2} - \frac{4r^2}{9} = \frac{r^2}{18} = \frac{1}{36a} \quad [\text{From (i)}]$$

Similarly, we shall get

$$E(Y) = \frac{2s}{3}, E(Y^2) = \frac{s^2}{2} \text{ and } \text{Var}(Y) = \frac{s^2}{18} = \frac{1}{36b}$$

Substituting in (v), we get

$$r(U, V) = \frac{1/(36a) - 1/(36b)}{1/(36a) + 1/(36b)} = \frac{b-a}{b+a}$$

**Example 10.13.** Let the random variable  $X$  have the marginal density

$$f_1(x) = 1, -\frac{1}{2} < x < \frac{1}{2}$$

and let the conditional density of  $Y$  be

$$\left. \begin{aligned} f(y|x) &= 1, x < y < x+1, -\frac{1}{2} < x < 0 \\ &= 1, -x < y < 1-x, 0 < x < \frac{1}{2} \end{aligned} \right\} \quad (*)$$

Show that the variables  $X$  and  $Y$  are uncorrelated.

**Solution.** We have

$$E(X) = \int_{-\frac{1}{2}}^{\frac{1}{2}} x f_1(x) dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x \cdot 1 dx = \left[ \frac{x^2}{2} \right]_{-\frac{1}{2}}^{\frac{1}{2}} = 0$$

If  $f(x, y)$  is the joint p.d.f. of  $X$  and  $Y$ , then

$$f(x, y) = f(y|x) f_1(x) = f(y|x). \quad (***) \quad [\because f_1(x) = 1]$$

$$\begin{aligned} E(XY) &= \int_{-\frac{1}{2}}^0 \int_x^{x+1} xy f(x, y) dx dy + \int_0^{\frac{1}{2}} \int_{-x}^{1-x} xy f(x, y) dx dy \\ &= \int_{-\frac{1}{2}}^0 \left[ x \int_x^{x+1} y dy \right] dx + \int_0^{\frac{1}{2}} \left[ x \int_{-x}^{1-x} y dy \right] dx \quad [\text{From (*) and}] \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{2} \int_{-\frac{1}{2}}^0 \dot{x}(2x+1)dx + \frac{1}{2} \int_0^{\frac{1}{2}} x(1-2x) dx \\
 &= \frac{1}{2} \left[ \frac{2}{3}x^3 + \frac{x^2}{2} \right]_0^{\frac{1}{2}} + \frac{1}{2} \left[ \frac{x^2}{2} - \frac{2}{3}x^3 \right]_0^{\frac{1}{2}} \\
 &= \frac{1}{2} \left[ \frac{1}{12} - \frac{1}{8} - \frac{1}{12} + \frac{1}{8} \right] = 0
 \end{aligned}$$

$$\therefore \text{Cov } (XY) = E(XY) - E(X) E(Y) = 0 \Rightarrow r(X, Y) = 0$$

Hence the variables  $X$  and  $Y$  are uncorrelated.

### EXERCISE 10(a)

1. (a) Show that the co-efficient of correlation  $r$  is independent of a change of scale and origin of the variables. Also prove that for two independent variables  $r = 0$ . Show by an example that the converse is not true. State the limits between which  $r$  lies and give its proof.

[Delhi Univ. M.Sc. (O.R.), 1986]

- (b) Let  $\rho$  be the correlation coefficient between two jointly distributed random variables  $X$  and  $Y$ . Show that  $|\rho| \leq 1$  and that  $|\rho| = 1$  if and only if  $X$  and  $Y$  are linearly related. [Indian Forest Service, 1991]

2. (a) Calculate the coefficient of correlation between  $X$  and  $Y$  for the following :

X...	1	3	4	5	7	8	10
Y...	2	6	8	10	14	16	20

Ans.  $r(X, Y) = +1$

(b) Discuss the statistical validity of the following statements :

- (i) "High positive coefficient of correlation between increase in the sale of newspapers and increase in the number of crimes leads to the conclusion that newspaper reading may be responsible for the increase in the number of crimes."

- (ii) "A high positive value of  $r$  between the increase in cigarette smoking and increase in lung cancer establishes that cigarette smoking is responsible for lung cancer."

- (c) (i) Do you agree with the statement that " $r = 0.8$  implies that 80% of the data are explained."

(ii) Comment on the following :

"The closeness of relationship between two variables is proportional to  $r$ ".

Hint. (a) No (b) Wrong.

- (d) By effecting suitable change of origin and scale, compute the product moment correlation coefficient for the following set of 5 observations on  $(X, Y)$  :

X :	-10	-5	0	5	10
Y :	5	9	7	11	13

Ans.  $r(X, Y) = 0.34$

3. The marks obtained by 10 students in Mathematics and Statistics are given below. Find the coefficient of correlation between the two subjects.

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics :	75	30	60	80	53	35	15	40	38	48
Marks in Statistics :	85	45	54	91	58	63	35	43	45	44

4. (a) The following table gives the number of blind per lakh of population in different age-groups. Find out the correlation between age and blindness.

Age in years	0—10	10—20	20—30	30—40	40—50
Number of blind per lakh	55	67	100	111	150
Age in year	50—60	60—70	70—80		
Number of blind per lakh	200	300.	500		

Ans. 0.89

(b) The following table gives the distribution of items of production and also the relatively defective items among them, according to size-groups. Is there any correlation between size and defect in quality?

Size-Group	15—16	16—17	17—18	18—19	19—20	20—21
No. of Items	200	270	340	360	400	300
No. of defective items	150	162	170	180	180	120

Hint. Here we have to find the correlation coefficient between the size-group ( $X$ ) and the percentage of defectives ( $Y$ ) given below.

$X$	15.5	16.5	17.5	18.5	19.5	20.5
$Y$	75	60	50	50	45	40

Ans.  $r = 0.94$ .

5. Using the formula

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2 r(X, Y) \sigma_X \sigma_Y$$

obtain the correlation coefficient between the heights of fathers ( $X$ ) and of the sons ( $Y$ ) from the following data:

$X$ :	65	66	67	68	69	70	71	67
$Y$ :	67	68	64	72	70	67	70	68

6. (a) From the following data, compute the co-efficient of correlation between  $X$  and  $Y$ .

	$X$ series	$Y$ series
No. of items	15	15
Arithmetic mean	25	18
Sum of squares of deviations from mean	136	138

Summation of product of deviations of  $X$  and  $Y$  series from the respective arithmetic means = 122.

Ans.  $r(X, Y) = 0.891$

(b) Coefficient of correlation between two variables  $X$  and  $Y$  is 0.32. Their covariance is 7.86. The variance of  $X$  is 10. Find the standard deviation of  $Y$  series.

(c) In two sets of variables  $X$  and  $Y$  with 50 observations each, the following data were observed :

$$\bar{X} = 10, \sigma_X = 3, \bar{Y} = 6, \sigma_Y = 2 \text{ and } r(X, Y) = 0.3$$

But on subsequent verification it was found that one value of  $X$  (= 10) and one value of  $Y$  (= 6) were inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of  $r$  affected ?

(Nagpur Univ. B.Sc., 1990)

Hint.  $\Sigma X = n\bar{X} = 500, \Sigma Y = n\bar{Y} = 300$

$$\Sigma X^2 = n(\sigma_X^2 + \bar{X}^2) = 5450, \Sigma Y^2 = 50(4 + 36) = 2000$$

$$r \sigma_X \sigma_Y = \text{Cov}(X, Y) = \frac{\Sigma XY}{n} - \bar{X} \bar{Y}$$

$$\Rightarrow 0.3 \times 3 \times 2 = \frac{\Sigma XY}{50} - 10 \times 6$$

$$\Rightarrow \Sigma XY = 50(1.8 + 60) = 3090$$

After weeding out the incorrect pair of observation, viz.,  $(X = 10, Y = 6)$ , the corrected values of  $\Sigma X$ ,  $\Sigma Y$ ,  $\Sigma X^2$ ,  $\Sigma Y^2$  and  $\Sigma XY$  for the remaining  $50 - 1 = 49$  pairs of observations are given below :

*Corrected Values :*

$$\Sigma X = 500 - 10 = 490; \Sigma Y = 300 - 6 = 294$$

$$\Sigma XY = 3090 - 10 \times 6 = 3090 - 60 = 3030$$

$$\Sigma X^2 = 5450 - 10^2 = 5350, \Sigma Y^2 = 2000 - 6^2 = 1964$$

$$\therefore r = \frac{\text{Corrected Cov}(X, Y)}{(\text{Corrected } \sigma_X) \times (\text{Corrected } \sigma_Y)} = \frac{90/49}{\sqrt{1450/49 \times 200/49}} = 0.3$$

Hence the correlation coefficient is invariant in this case.

(d) A prognostic test in Mathematics was given to 10 students who were about to begin a course in Statistics. The scores ( $X$ ) in their test were examined in relation to scores ( $Y$ ) in the final examination in Statistics. The following results were obtained :—

$$\Sigma X = 71, \Sigma Y = 70, \Sigma X^2 = 555, \Sigma Y^2 = 526 \text{ and } \Sigma XY = 527$$

Find the coefficient of correlation between  $X$  and  $Y$ .

(Kerala Univ. B.Sc., 1990)

7. (a)  $X_1$  and  $X_2$  are independent variables with means 5 and 10 and standard deviations 2 and 3 respectively. Obtain  $r(U, V)$  where

$$U = 3X_1 + 4X_2 \text{ and } V = 3X_1 - X_2$$

Ans. 0

(Delhi Univ. B.Sc., 1988)

(b) If  $X$  and  $Y$  are normal and independent with zero means and standard deviations 9 and 12 respectively, and if  $X + 2Y$  and  $kX - Y$  are non-correlated, find  $k$ .

(c)  $X, Y, Z$  are random variables each with expectation 10 and variances 1, 4 and 9 respectively. The correlation coefficients are

$$r(X, Y) = 0, r(Y, Z) = r(X, Y) = 1/4$$

Obtain the numerical values of :

- (i)  $E(X + Y - 2Z)$ , (ii)  $\text{Cov}(X + 3, Y + 3)$ , (iii)  $V(X - 2Z)$  and  
(iv)  $\text{Cov}(3X, 5Z)$

**Ans.** (i) 0, (ii) 0, (iii) 34, and (iv) 45/4.

(d)  $X$  and  $Y$  are discrete random variables. If  $\text{Var}(X) = \text{Var}(Y) = \sigma^2$ ,

$\text{Cov}(X, Y) = \frac{\sigma^2}{2}$ , find (i)  $\text{Var}(2X - 3Y)$ , (ii)  $\text{Corr}(2X + 3, 2Y - 3)$ .

8. (a) Prove that :

$$V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2ab \text{Cov}(X, Y)$$

Hence deduce that if  $X$  and  $Y$  are independent

$$V(X \pm Y) = V(X) + V(Y)$$

(b) Prove that correlation coefficient between  $X$  and  $Y$  is positive or negative according as

$$\sigma_{X+Y} > \text{or} < \sigma_{X-Y}$$

9. Show that if  $X$  and  $Y$  are two random variables each assuming only two values and the correlation co-efficient between them is zero, then they are independent. Indicate with justification whether the result is true in general.

Find the correlation coefficient between  $X$  and  $a - X$ , where  $X$  is any random variable and  $a$  is constant.

10. (a)  $X_i$  ( $i = 1, 2, 3$ ) are uncorrelated variables each having the same standard deviation. Obtain the correlation between  $X_1 + X_2$  and  $X_2 + X_3$ .

**Ans.** 1/2

(b) If  $X_i$  ( $i = 1, 2, 3$ ) are three uncorrelated variables having standard deviations  $\sigma_1, \sigma_2$  and  $\sigma_3$  respectively, obtain the coefficient of correlation between  $(X_1 + X_2)$  and  $(X_2 + X_3)$ .

**Ans.**  $\sigma_2^2 / \sqrt{(\sigma_1^2 + \sigma_2^2)(\sigma_2^2 + \sigma_3^2)}$

(c) Two random variables  $X$  and  $Y$  have zero means, the same variance  $\sigma^2$  and zero correlation. Show that

$$U = X \cos \alpha + Y \sin \alpha \quad \text{and} \quad V = X \sin \alpha - Y \cos \alpha$$

have the same variance  $\sigma^2$  and zero correlation.

*(Bangalore Univ. B.Sc., 1991)*

(d) Let  $X$  and  $Y$  be uncorrelated random variables. If  $U = X + Y$  and  $V = X - Y$ , prove that the coefficient of correlation between  $U$  and  $V$  is  $(\sigma_X^2 - \sigma_Y^2) / (\sigma_X^2 + \sigma_Y^2)$ , where  $\sigma_X^2$  and  $\sigma_Y^2$  are variances of  $X$  and  $Y$  respectively.

(e) Two independent random variables  $X$  and  $Y$  have the following variances :  $\sigma_X^2 = 36, \sigma_Y^2 = 16$ . Calculate the coefficient of correlation between

$$U = X + Y \text{ and } V = X - Y$$

(f) Random variables  $X$  and  $Y$  have zero means and non-zero variances  $\sigma_X^2$  and  $\sigma_Y^2$ . If  $Z = Y - X$ , then find  $\sigma_Z$  and the correlation coefficient  $\rho(X, Z)$  of  $X$  and  $Z$  in terms of  $\sigma_X$ ,  $\sigma_Y$  and the correlation coefficient  $\rho(X, Y)$  of  $X$  and  $Y$ .

(g) If the independent random variables  $X_1, X_2$  and  $X_3$  have the means 4, 9 and 3 and variances 3, 7, 5, respectively, obtain the mean and variance of

$$(i) \quad Y = 2X_1 - 3X_2 + 4X_3, \quad (ii) \quad Z = X_1 + 2X_2 - X_3, \text{ and}$$

(iii) Calculate the correlation between  $Y$  and  $Z$ .

[Delhi Univ. M.A.(Eco.), 1989]

11. (a)  $X_1, X_2, \dots, X_n$  are uncorrelated random variables, all with the same distribution and zero means. Let  $\bar{X} = \sum X_i/n$

Find the correlation coefficient between (i)  $X_i$  and  $\bar{X}$  and (ii)  $X_i - \bar{X}$  and  $\bar{X}$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1993]

Hint.  $r(X_i, \bar{X}) = \frac{\sigma^2/n}{\sqrt{\sigma^2 \cdot \sigma^2/n}} = \frac{1}{\sqrt{n}}$

$$\begin{aligned} \text{Cov}(X_i - \bar{X}, \bar{X}) &= \text{Cov}(X_i, \bar{X}) - \text{Var}(\bar{X}) \\ &= (\sigma^2/n) - (\sigma^2/n) = 0 \end{aligned}$$

$$\therefore r(X_i - \bar{X}, \bar{X}) = 0$$

(b)  $X_1, X_2, \dots, X_n$  are random variables each with the same expected value  $\mu$  and s.d.  $\sigma$ . The correlation coefficient between any two  $X$ 's is  $\rho$ . Show

that (i)  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n} + \left(1 - \frac{1}{n}\right)\rho\sigma^2$ ,

$$(ii) \quad E \sum_{i=1}^n (X_i - \bar{X})^2 = (n-1)(1-\rho)\sigma^2, \text{ and } (iii) \quad \rho > -\frac{1}{n-1}$$

12. (a) If  $X$  and  $Y$  are independent random variables, show that

$$r(X+Y, X-Y) = r^2(X, X+Y) - r^2(Y, X+Y),$$

where  $r(X+Y, X-Y)$  denotes the coefficient of correlation between  $(X+Y)$  and  $(X-Y)$ .

(Meerut Univ. B.Sc., 1991)

(b) Let  $X$  and  $Y$  be random variables having mean 0, variance 1 and correlation  $r$ . Show that  $X - rY$  and  $Y$  are uncorrelated and that  $X - rY$  has mean zero and variance  $1 - r^2$ .

13.  $X_1$  and  $X_2$  are two variables with zero means, variances  $\sigma_1^2$  and  $\sigma_2^2$  respectively and  $r$  is the correlation coefficient between them. Determine the values of the constants  $a$  and  $b$  which are independent of  $r$  such that  $X_1 + aX_2$  and  $X_1 + bX_2$  are uncorrelated.

14. (a) If  $X_1$  and  $X_2$  are two random variables with means  $\mu_1$  and  $\mu_2$ , variances  $\sigma_1^2, \sigma_2^2$  and correlation coefficient  $r$ , find the correlation coefficient between

$$U = a_1X_1 + a_2X_2 \text{ and } V = b_1X_1 + b_2X_2,$$

where  $a_1, a_2$  and  $b_1, b_2$  are constants.

(b) Let  $X_1, X_2$  be independent random variables with means  $\mu_1, \mu_2$  and non-zero variances  $\sigma_1^2, \sigma_2^2$  respectively. Let  $U = X_1 - X_2$  and  $V = X_1 X_2$ . Find the correlation coefficient between (i)  $X_1$  and  $U$ , (ii)  $X_1$  and  $V$ , in terms of  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ .

15. (a) If  $U = aX + bY$  and  $V = bX - aY$ , where  $X$  and  $Y$  are measured from their respective means and if  $U$  and  $V$  are uncorrelated,  $r$  the co-efficient of correlation between  $X$  and  $Y$  is given by the equation.

$$\sigma_U \sigma_V = (a^2 + b^2) \sigma_X \sigma_Y (1 - r^2)^{1/2} \quad (\text{Utkal Univ. B. Sc., 1993})$$

(b) Let  $U = aX + bY$  and  $V = aX - bY$  where  $X, Y$  represent deviations from the means of two measurements on the same individual. The coefficient of correlation between  $X$  and  $Y$  is  $\rho$ . If  $U, V$  are uncorrelated, show that

$$\sigma_U \sigma_V = 2ab\sigma_X \sigma_Y (1 - r^2)^{1/2}$$

16. Show that, if  $a$  and  $b$  are constants and  $r$  is the correlation coefficient between  $X$  and  $Y$ , then the correlation coefficient between  $aX$  and  $bY$  is equal to  $r$  if the signs of  $a$  and  $b$  are alike, and to  $-r$  if they are different.

Also show that, if constants  $a, b$  and  $c$  are positive, the correlation coefficient between  $(aX + bY)$  and  $cY$  is equal to

$$(ar\sigma_X + b\sigma_Y) / \sqrt{(a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_X\sigma_Y)}$$

17. If  $X_1, X_2$  and  $X_3$  are three random variables measured from their respective means as origin and of equal variances, find the coefficient of correlation between  $X_1 + X_2$  and  $X_2 + X_3$  in terms of  $r_{12}, r_{13}$  and  $r_{23}$  and show that it is equal to

$$(i) \frac{r_{12} + 1}{2}, \text{ if } r_{13} = r_{23} = 0, \text{ and } (ii) \frac{r_{12} + 3}{4}, \text{ if } r_{13} = r_{23} = 1$$

18. (a) For a weighted distribution  $(x_i, w_i)$ , ( $i = 1, 2, \dots, n$ ) show that the weighted arithmetic mean  $\bar{x}_w = \sum w_i x_i / \sum w_i >$  or  $<$  the unweighted mean  $\bar{x} = \sum x_i / n$  according as  $r_{xw} >$  or  $< 0$ .

(b) Given  $N$  values  $x_1, x_2, \dots, x_N$  of variable  $X$  and weights  $w_1, w_2, \dots, w_N$ , express the coefficient of correlation between  $X$  and  $W$  in terms involving the difference between the arithmetic mean and the weighted mean of  $X$ .

19. (a) A coin is tossed  $n$  times. If  $X$  and  $Y$  denote the (random) number of heads and number of tails turned up respectively, show that  $r(X, Y) = -1$ .

**Hint.** Note that  $X + Y = n \Rightarrow Y = n - X$

$$\therefore r(X, Y) = r(X, n - X) = r(X, -X) = -r(X, X) = -1.$$

(b) Two dice are thrown, their scores being  $a$  and  $b$ . The first die is left on the table while the second is picked up and thrown again giving the score  $c$ . Suppose the process is repeated a large number of times. What is the correlation coefficient between  $X = a + b$  and  $Y = a + c$ ?

$$\text{Ans. } r(X, Y) = \frac{1}{2}$$

20. (a) If  $X$  and  $Y$  are independent random variables with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2, \sigma_2^2$  respectively, show that the correlation coefficient between  $U = X$  and  $V = X - Y$  in terms of  $\mu_1, \mu_2, \sigma_1^2$  and  $\sigma_2^2$  is  $\sigma_1 / \sqrt{\sigma_1^2 + \sigma_2^2}$ .

(b) If  $X$  and  $Y$  are independent random variables with non-zero variances, show that the correlation coefficient between  $U = XY$  and  $V = X$  in terms of mean and variance of  $X$  and  $Y$  is given by

$$\mu_2\sigma_1/\sqrt{\sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}$$

[Delhi Univ. B.Sc. (Stat Hons.), 1987]

21. If  $X_i$ ,  $Y_j$  and  $Z_k$  are all independent random variables with mean zero and unit variance, find the correlation coefficient between

$$U = \sum_{i=1}^m X_i + \sum_{j=1}^n Y_j \text{ and } V = \sum_{i=1}^m X_i + \sum_{k=1}^n Z_k$$

Ans.  $r(U, V) = m/(m + n)$

(Bombay Univ., B.Sc, 1990)

22. (a) Find the value of  $l$  so that the correlation coefficient between  $(X - lY)$  and  $(X + Y)$  is maximum, where  $X, Y$  are independent random variables each with mean zero and variance 1. [Ans.  $l = -1$ ]

Hint.  $U = X - lY$ ;  $V = X + Y$ . Now find  $l$  so that  $r(U, V) = 1$ .

(b) If  $U = X + kY$  and  $V = X + mY$  and  $r$  is the correlation coefficient between  $X$  and  $Y$ , find the correlation coefficient between  $U$  and  $V$ . Show that

$U$  and  $V$  are uncorrelated if  $k = \frac{-\sigma_X(\sigma_X + rm\sigma_Y)}{\sigma_Y(r\sigma_X + m\sigma_Y)}$

and further if  $m = \frac{\sigma_X}{\sigma_Y}$ , then  $k = -\frac{\sigma_X}{\sigma_Y}$ . (Gujarat Univ. M.A., 1993)

23.  $X_1, X_2, X_3$  are three variables, each with variance  $\sigma^2$  and the correlation coefficient between any two of them is  $r$ . If  $\bar{X} = (X_1 + X_2 + X_3)/3$ , show that

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{3}(1 + 2r)$$

Deduce that  $r \geq -1/2$ .

24. (a) If  $U = aX + bY$  and  $V = bX - aY$ , show that  $U$  and  $V$  are uncorrelated if  $\frac{ab}{a^2 - b^2} = \frac{\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}$

where  $\rho$  is the coefficient of correlation between  $X$  and  $Y$ . Show further that, in this case

$$\sigma_U^2 + \sigma_V^2 = (a^2 + b^2)(\sigma_X^2 + \sigma_Y^2) \text{ and } \sigma_U\sigma_V = (a^2 + b^2)\sigma_X\sigma_Y\sqrt{1 - \rho^2}$$

(b) If  $u = aX + bY$ ,  $v = cX + dY$ , show that

$$\begin{vmatrix} \text{var}(u) & \text{cov}(u, v) \\ \text{cov}(u, v) & \text{var}(v) \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix}^2 \begin{vmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(X, Y) & \text{var}(Y) \end{vmatrix}$$

25. If  $X$  is a standard normal variate and  $Y = a + bX + cX^2$ ,

where  $a, b, c$  are constants, find the correlation coefficient between  $X$  and  $Y$ . Hence or otherwise obtain the conditions when (i)  $X$  and  $Y$  are uncorrelated and (ii)  $X$  and  $Y$  are perfectly correlated.

26. (a) If  $X \sim N(0, 1)$ , find  $\text{corr}(X, Y)$  where  $Y = a + bX + cX^2$ .

[Delhi Univ. B.Sc. (Maths. Hons.), 1985]

$$\text{Ans. } r(X, Y) = \frac{b}{\sqrt{b^2 + 2c^2}}$$

- (b) If  $X$  has Laplace distribution with parameters  $(\lambda, 0)$  and  $Y = a + bX + cX^2$ , find  $r(X, Y)$

[Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1989]

$$\text{Hint. } p(x) = \frac{1}{2} \lambda \exp[-\lambda |x|], -\infty < x < \infty.$$

$$E(X^{2k+1}) = 0 = \mu_{2k+1}; E(X^{2k}) = \mu_{2k} = (2k)! / \lambda^{2k}$$

$$r_{XY} = \frac{\lambda b}{\sqrt{b^2 \lambda^2 + 10c^2}}$$

27. In a sample of  $n$  random observations from exponential distribution with parameter  $\lambda$ , the number of observations in  $(0, 1/\lambda)$  and  $(1/\lambda, 2/\lambda)$ , denoted by  $X$  and  $Y$  are noted. Find  $r(X, Y)$ .

$$\text{Hint. } p_1 = p(0 < X < 1/\lambda) = \int_0^{1/\lambda} \lambda e^{-\lambda x} dx = \frac{e - 1}{e}$$

$$p_2 = p(1/\lambda < Y < 2/\lambda) = \int_{1/\lambda}^{2/\lambda} \lambda e^{-\lambda y} dy = \frac{e - 1}{e^2}$$

- Then  $(X, Y)$  has a trinomial distribution with parameters  $(n = 3, p_1, p_2, p_3 = 1 - p_1 - p_2)$ .

Hence we have

$$r(X, Y) = - \left[ \frac{p_1 p_2}{(1 - p_1)(1 - p_2)} \right]^{1/2} = - \frac{e - 1}{\sqrt{e^2 - e + 1}}.$$

28. Prove that :

$$r(X, Y+Z) = \frac{\sigma_Y}{\sigma_{Y+Z}} \cdot r(X, Y) + \frac{\sigma_Z}{\sigma_{Y+Z}} \cdot r(X, Z)$$

29. If  $X$  and  $Y$  are independent random variables, find  $\text{Corr}(X, XY)$ . Deduce the value of  $\text{Corr}(X, X/Y)$ .

$$\text{Ans. } r(X, XY) = \sigma_X \mu_Y / [\sigma_X^2 \sigma_Y^2 + \mu_X^2 \sigma_Y^2 + \mu_Y^2 \sigma_X^2]^{1/2}$$

30. Prove or Disprove :

$$(a) r(X, Y) = 0 \Rightarrow r(|X|, Y) = 0$$

$$(b) r(X, Y) = 0, r(Y, Z) = 0 \Rightarrow r(X, Z) = 0.$$

**Ans.** (a) False, unless  $X$  and  $Y$  are independent.

(b) Hint. Let  $Z = X$ , and  $X$  and  $Y$  be independent. Then

$$r(X, Y) = 0 = r(Y, Z). \text{ But } r(X, Z) = r(X, X) = 1.$$

31. Let random variable  $X$  have a p.d.f.  $f(\cdot)$  with distribution function  $F(\cdot)$ , mean  $\mu$  and variance  $\sigma^2$ . Define  $Y = \alpha + \beta X$ , where  $\alpha$  and  $\beta$  are constants satisfying  $-\infty < \alpha < \infty$ , and  $\beta > 0$ .

(a) Select  $\alpha$  and  $\beta$  so that  $Y$  has mean 0 and variance 1.

(b) What is the correlation coefficient between  $X$  and  $Y$ ?

32. Let  $(X, Y)$  be jointly discrete random variables such that each  $X$  and  $Y$  have at most two mass points. Prove or disprove :  $X$  and  $Y$  are independent if and only if they are uncorrelated:

Ans. True.

33. If the variables  $X_1, X_2, \dots, X_{2n}$  all have the same variance  $\sigma^2$  and the correlation coefficient between  $X_i$  and  $X_j$  ( $i \neq j$ ) has the same value, show that the correlation between  $\sum_{i=1}^n X_i$  and  $\sum_{j=n+1}^{2n} X_j$  is given by  $[np/(1 + (n-1)\rho)]$ .

34. The means of independent r.v's  $X_1, X_2, \dots, X_n$  are zero and variances are equal, say unity. The correlation coefficients between the sum of selected  $t$  ( $< n$ ) variables out of these variables and the sum of all  $n$  variables are found out. Prove that the sum of squares of all these correlation coefficients is  ${}^{n-1}C_{t-1}$ .

[Burdwan Univ. B.Sc. (Hons.), 1989]

35. Two variables  $U$  and  $V$  are made up of the sum of a number of terms as follows :

$$U = X_1 + X_2 + \dots + X_a + Y_1 + Y_2 + \dots + Y_b,$$

$$V = X_1 + X_2 + \dots + X_a + Z_1 + Z_2 + \dots + Z_b,$$

where  $a$  and  $b$  are all suffixes and where  $X$ 's,  $Y$ 's and  $Z$ 's are all uncorrelated standardised random variables. Show that the correlation coefficient between

$U$  and  $V$  is  $\frac{n}{\sqrt{(n+a)(n+b)}}$ . Show further that

$$\begin{aligned} \xi &= \sqrt{(n+b)} U + \sqrt{(n+a)} V \\ \eta &= \sqrt{(n+b)} U - \sqrt{(n+a)} V \end{aligned} \quad \dots (*)$$

are uncorrelated

[South Gujarat Univ. B.Sc., 1989]

36. (a) Let the random variables  $X$  and  $Y$  have the joint p.d.f.

$$f(x, y) = 1/3 ; (x, y) = (0, 0), (1, 1) (2, 0)$$

Compute  $E(X)$ ,  $V(X)$ ,  $E(Y)$ ,  $V(Y)$  and  $r(X, Y)$ . Are  $X$  and  $Y$  stochastically independent ? Give reasons.

(b) Let  $(X, Y)$  have the probability distribution :

$$f(0, 0) = 0.45, f(0, 1) = 0.05, f(1, 0) = 0.35, f(1, 1) = 0.15.$$

Evaluate  $V(X)$ ,  $V(Y)$  and  $r(X, Y)$ .

Show that while  $X$  and  $Y$  are correlated,  $X$  and  $X-5Y$  are uncorrelated. Are  $X$  and  $X-5Y$  independent ?

(c) Given the bivariate probability distribution :

$$f(-1, 0) = 1/15, \quad f(-1, 1) = 3/15, \quad f(-1, 2) = 2/15$$

$$f(0, 0) = 2/15, \quad f(0, 1) = 2/15, \quad f(0, 2) = 1/15$$

$$f(1, 0) = 1/15, \quad f(1, 1) = 1/15, \quad f(1, 2) = 2/15$$

$$f(x, y) = 0, \text{ elsewhere.}$$

Obtain :

(i) The marginal distributions of  $X$  and  $Y$ .

- (ii) The conditional distributions of  $Y$  given  $X = 0$ .
- (iii)  $E(Y|X = 0)$ .
- (iv) The product moment correlation coefficient between  $X$  and  $Y$ .  
Are  $X$  and  $Y$  independently distributed?

37. If  $X$  and  $Y$  are standardised variates with correlation coefficient  $\rho$ , prove that  $E[\max(X^2, Y^2)] \leq 1 + \sqrt{1 - \rho^2}$

**Hint.**  $\max(X^2, Y^2) = \frac{1}{2}|X^2 - Y^2| + \frac{1}{2}(X^2 + Y^2)$  ...(\*)

$$E(X) = E(Y) = 0; E(X^2) = E(Y^2) = 1; E(XY) = \rho$$

and  $[E|X - Y|, |X + Y|]^2 \leq E(X - Y)^2 \cdot E(X + Y)^2$

(By Cauchy-Schwartz Inequality)

38. The joint p.d.f. of two variates  $X$  and  $Y$  is given by

$$f(x, y) = k[(x+y) - (x^2 + y^2)]; 0 < (x, y) < 1$$

= 0, otherwise.

Show that  $X$  and  $Y$  are uncorrelated but not independent.

39(a). If the random variables  $X$  and  $Y$  have the joint p.d.f.,

$$f(x, y) = \begin{cases} x + y; & 0 < x < 1, 0 < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

then show that the correlation coefficient between  $X$  and  $Y$  is  $-\frac{1}{11}$ .

[Madras Univ. B.Sc., Oct., 1990]

(b) The density function  $f$  of a random variable  $X$  is given by

$$f(x) = \begin{cases} kx^2, & \text{if } -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

- (i) What is the value of  $k$ ? What is the distribution function of  $X$ ?
- (ii) Obtain the density function of the random variable  $Y = X^2$ .
- (iii) Obtain the correlation coefficient between  $X$  and  $Y$ .
- (iv) Are  $X$  and  $Y$  independently distributed?

40(a). If  $f(x, y) = \frac{6-x-y}{8}; 0 \leq x \leq 2, 2 \leq y \leq 4$ ,

find (i)  $\text{Var}(X)$ , (ii)  $\text{Var}(Y)$  (iii)  $r(X, Y)$ .

**Ans.** (i)  $\frac{11}{36}$ , (ii)  $\frac{11}{36}$ , (iii)  $-\frac{1}{11}$ .

(b) Given the joint density of random variables  $X, Y, Z$  as :

$$f(x, y, z) = k x \exp[-(y+z)], 0 < x < 2, y \geq 0, z \geq 0$$

= 0, elsewhere

Find

- (i)  $k$ ,
- (ii) the marginal density function,
- (iii) conditional expectation of  $Y$ , given  $X$  and  $Z$ , and
- (iv) the product moment correlation between  $X$  and  $Y$ .

[Madras Univ. B.Sc. (Main Stat.), 1988]

(c) Suppose that the two dimensional random variable  $(X, Y)$  has p.d.f. given by  $f(x, y) = ke^{-y}$ ,  $0 < x < y < 1$   
 $= 0$ , elsewhere

Find the correlation coefficient  $r_{XY}$ . [Delhi Univ. M.C.A., 1991]

41. The joint density of  $(X, Y)$  is :

$$f(x, y) = \frac{1}{8}(x + y), \quad 0 \leq x \leq 2, 0 \leq y \leq 2.$$

Find  $\mu'_{rs} = E(X^r Y^s)$  and hence find  $\text{Corr}(X, Y)$ .

$$\text{Ans. } \mu'_{rs} = 2^{r+s} \left[ \frac{1}{(r+2)(s+1)} + \frac{1}{(r+1)(s+2)} \right]; r = -\frac{1}{11}.$$

(b) Find the m.g.f. of the bivariate distribution :

$$\begin{aligned} f(x, y) &= 1, \quad 0 < (x, y) < 1 \\ &= 0, \text{ otherwise} \end{aligned}$$

and hence find  $r(X, Y)$ .

$$\text{Ans. } M(t_1, t_2) = (e^{t_1} - 1)(e^{t_2} - 1)/(t_1 t_2); t_1 \neq 0, t_2 \neq 0, r(X, Y) = 0.$$

42. Let  $(X, Y)$  have joint density :

$$f(x, y) = e^{-(x+y)} I_{(0, \infty)}(x) \cdot I_{(0, \infty)}(y)$$

Find  $\text{Corr}(X, Y)$ . Are  $X$  and  $Y$  independent?

Ans.  $\text{Corr}(X, Y) = 0$ :  $X$  and  $Y$  are independent.

43. A bivariate distribution in two discrete random variables  $X$  and  $Y$  is defined by the probability generating function :

$$\exp[a(u-1) + b(v-1) + c(u-1)(v-1)],$$

simultaneous probability of  $X = r \cap Y = s$ , where  $r$  and  $s$  are integers being the coefficient of  $u^r v^s$ . Find the correlation coefficient between  $X$  and  $Y$ .

Hint. Put  $u = e^{t_1}$  and  $v = e^{t_2}$  in  $\exp[a(u-1) + b(v-1) + c(u-1)(v-1)]$ , the result will be the m.g.f. of a bivariate distribution and is given by

$$M(t_1, t_2) = \exp[a(e^{t_1} - 1) + b(e^{t_2} - 1) + c(e^{t_1} - 1)(e^{t_2} - 1)]$$

$$\text{We have } \left[ \frac{\partial M}{\partial t_1} \right]_{t_1=t_2=0} = a, \quad \left[ \frac{\partial^2 M}{\partial t_1^2} \right]_{t_1=t_2=0} = a(a+1).$$

$$\left[ \frac{\partial^2 M}{\partial t_1 \partial t_2} \right]_{\substack{t_1=0 \\ t_2=0}} = ab + c, \quad \left[ \frac{\partial M}{\partial t_2} \right]_{\substack{t_1=0 \\ t_2=0}} = b, \quad \left[ \frac{\partial^2 M}{\partial t_2^2} \right]_{\substack{t_1=0 \\ t_2=0}} = b(b+1)$$

So we have

$$E(X) = a, E(X^2) = a(a+1), E(Y) = b, E(Y^2) = b(b+1) \text{ and } E(XY) = ab + c$$

$$\therefore r(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{[E(X^2) - \{E(X)\}^2][E(Y^2) - \{E(Y)\}^2]}} = \frac{c}{\sqrt{ab}}$$

44. Let the number  $X$  be chosen at random from among the integers 1, 2, 3, 4 and the number  $Y$  be chosen from among those at least as large as  $X$ . Prove that  $\text{Cov}(X, Y) = 5/8$ . Find also the regression line of  $Y$  on  $X$ .

[Delhi Univ. B.Sc. (Maths. Hons.), 1990]

Hint.  $P(X = k) = \frac{1}{4}; k = 1, 2, 3, 4$  and  $Y \geq X$ .

$$\hat{P}(Y=y | X=1) = \frac{1}{4}; y = 1, 2, 3, 4 (\because y \geq x);$$

$$P(Y=y | X=2) = \frac{1}{3}, y = 2, 3, 4$$

$$P(Y=y | X=3) = \frac{1}{2}, y = 3, 4; P(Y=y | X=4) = 1, y = 4.$$

The joint probability distribution can be obtained on using :

$$P(X=x, Y=y) = P(X=x) \cdot P(Y=y | X=x).$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{5/8}{\sqrt{(5/4) \times (41/48)}} = \sqrt{\frac{15}{41}}$$

$$\text{Regression line of } Y \text{ on } X : Y - E(Y) = \frac{r \sigma_Y}{\sigma_X} [X - E(X)]$$

**45.** Two ideal dice are thrown. Let  $X_1$  be the score on the first dice and  $X_2$ , the score on the second dice. Let  $Y = \max\{X_1, X_2\}$ . Obtain the joint distribution of  $Y$  and  $X_1$  and show that

$$\text{Corr}((Y, X_1)) = \frac{3}{2 \sqrt{73}}$$

**46.** Consider an experiment of tossing two tetrahedra. Let  $X$  be the number of the down turned face of first tetrahedron and  $Y$ , the larger of the two numbers. Obtain the joint distribution of  $X$  and  $Y$  and hence  $\rho(X, Y)$ .

$$\text{Ans. } \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{5/8}{\sqrt{5/4} \cdot \sqrt{55/64}} = \frac{2}{\sqrt{11}}$$

**47.** Three fair coins are tossed. Let  $X$  denote the number of heads on the first two coins and let  $Y$  denote the number of tails on the last two coins.

(a) Find the joint distribution of  $X$  and  $Y$ .

(b) Find the conditional distribution of  $Y$  given that  $X=1$ .

(c) Find  $\text{Cov}(X, Y)$ .

$$\text{Ans. Cov}(X, Y) = -1/4.$$

**48.** For the trinomial distribution of two random variables  $X$  and  $Y$ :

$$f(x, y) = \frac{n!}{x! y! (n-x-y)!} p^x q^y (1-p-q)^{n-x-y}$$

for  $x, y = 0, 1, 2, \dots, n$  and  $x+y \leq n$ ,  $p \geq 0$ ,  $q \geq 0$  and  $p+q \leq 1$ .

(a) Obtain the marginal distribution of  $Y$

(b) Obtain  $E(X|Y=y)$ .

(c) Find  $\rho(X, Y)$ .

$$\text{Ans. (a) } X \sim B(n, p), Y \sim B(n, q)$$

$$(b) (X | Y=y) \sim B\left(n-y, \frac{p}{1-q}\right)$$

(Note :  $p+q \neq 1$ )

$$\therefore E(X | Y=y) = (n-y) \left( \frac{p}{1-q} \right)$$

$$(c) \text{Cov}(X, Y) = -npq; \rho(X, Y) = -\left[\frac{pq}{(1-p)(1-q)}\right]^{1/2}$$

### OBJECTIVE TYPE QUESTIONS

I. Comment on the following :

- (i)  $r_{XY} = 0 \Rightarrow X$  and  $Y$  are independent.
- (ii) If  $r_{XY} > 0$  then  $r_{X^2, Y} > 0$ ,  $r_{X^2, Y^2} > 0$  and  $r_{X^2, Y^2} > 0$
- (iii)  $r_{XY} > 0 \Rightarrow E(XY) > E(X)E(Y)$
- (iv) Pearson's coefficient of correlation is independent of origin but not of scale.
- (v) The numerical value of product moment correlation coefficient ' $r$ ' between two variables  $X$  and  $Y$  cannot exceed unity.
- (vi) If the correlation coefficient between the variables  $X$  and  $Y$  is zero then the correlation coefficient between  $X^2$  and  $Y^2$  is also zero.
- (vii) If  $r > 0$ , then as  $X$  increases,  $Y$  also increases.
- (viii) "The closeness of relationship between two variables is proportional to  $r$ ."
- (ix)  $r$  measures every type of relationship between the two variables.

II. Comment on the following values of ' $r$ ' (correlation coefficient) :

1, -0.95, 0, -1.64, 0.87, 0.32, -1, 2.4.

III. (i) If  $\rho_{XY} = -0.9$ , then for large values of  $X$ , what sort of values do we expect for  $Y$ ?

(ii) If  $\rho_{XY} = 0$ , what is the value of  $\text{cov}(X, Y)$  and how are  $X$  and  $Y$  related?

IV. Indicate the correct answer :

- (i) The coefficient of correlation will have positive sign when
  - (a)  $X$  is increasing,  $Y$  is decreasing, (b) both  $X$  and  $Y$  are increasing,
  - (c)  $X$  is decreasing,  $Y$  is increasing, (d) there is no change in  $X$  and  $Y$ .
- (ii) The coefficient of correlation (a) can take any value between -1 and +1
  - (b) is always less than -1, (c) is always more than +1, (d) cannot be zero.
- (iii) The coefficient of correlation (a) cannot be positive, (b) cannot be negative, (c) is always positive, (d) can be both positive as well as negative.

(iv) Probable error of  $r$  is

$$(a) 0.6475 \frac{1-r^2}{\sqrt{n}}, (b) 0.6754 \frac{1+r^2}{\sqrt{n}}, (c) 0.6547 \frac{1-r^2}{n},$$

$$(d) 0.6754 \frac{1-r^2}{n}.$$

(v) The coefficient of correlation between  $X$  and  $Y$  is 0.6. Their covariance is 4.8. The variance of  $X$  is 9. Then the S.D. of  $Y$  is

$$(a) \frac{4.8}{3 \times 0.6}, (b) \frac{0.6}{4.8 \times 3}, (c) \frac{3}{4.8 \times 0.6}, (d) \frac{4.8}{9 \times 0.6}.$$

- (vi) The coefficient of correlation is independent of (a) change of scale only, (b) change of origin only, (c) both change of scale and origin, (d) neither change of scale nor change of origin.

V. Fill in the blanks :

- The Karl Pearson coefficient of correlation between variables  $X$  and  $Y$  is ... ...
- Two independent variables are ... ...
- Limits for correlation coefficient are ... ...
- If  $r$  be the correlation coefficient between the random variables  $X$  and  $Y$  then the variance of  $X + Y$  is ... ...
- The absolute value of the product moment correlation coefficient is less than ... ...
- Correlation coefficient is invariant under changes of ... and ... ...

VI. How can you use scatter diagram to obtain an idea of extent and nature (direction) of the correlation coefficient ?

**10-4. Calculation of the Correlation Coefficient for a Bivariate Frequency Distribution.** When the data are considerably large, they may be summarised by using a two-way table. Here, for each variable a suitable number of classes are taken, keeping in view the same considerations as in the univariate case. If there are  $n$  classes for  $X$  and  $m$  classes for  $Y$ , there will be in all  $m \times n$  cells in the two-way table. By going through the pairs of values of  $X$  and  $Y$ , we can find the frequency for each cell. The whole set of cell frequencies will then define a *bivariate frequency distribution*. The column totals and row totals will give us the marginal distributions of  $X$  and  $Y$ . A particular column or row will be called the conditional distribution of  $Y$  for given  $X$  or of  $X$  for given  $Y$  respectively.

Suppose that the bivariate data on  $X$  and  $Y$  are presented in a two-way correlation table (shown on page 10-33) where there are  $m$  classes of  $Y$  placed along the horizontal line and  $n$  classes of  $X$  along a vertical line and  $f_{ij}$  is the frequency of individuals lying in the  $(i, j)$ th cell.

Here

$$\sum_x f(x, y) = g(y)$$

is the sum of the frequencies along any row and

$$\sum_y f(x, y) = f(x)$$

is the sum of the frequencies along any column. We observe that

$$\text{Thus } \sum_x \sum_y f(x, y) = \sum_y \sum_x f(x, y) = \sum_x f(x) = \sum_y g(y) = N$$

$$\bar{x} = \frac{1}{N} \sum_x \sum_y x f(x, y) = \frac{1}{N} \left[ \sum_x \left\{ x \sum_y f(x, y) \right\} \right] = \frac{1}{N} \sum_x x f(x)$$

$$\text{Similarly } \bar{y} = \frac{1}{N} \sum_y \sum_x y f(x, y) = \frac{1}{N} \sum_y y g(y)$$

$$\sigma_x^2 = \frac{1}{N} \sum_x \sum_y x^2 f(x, y) - \bar{x}^2 = \frac{1}{N} \sum_x x^2 f(x) - \bar{x}^2$$

## BIVARIATE FREQUENCY TABLE (CORRELATION TABLE)

	$X$ Series →	Classes					$\text{Total of frequencies of } Y \\ g(y)$
		$x_1$	$x_2$	$\dots x_i, \dots$	$\dots$	$x_m$	
$Y$ Series ↓	$y_1$						$\sum_x f(x, y) = g(y)$
	$y_2$						
	$\vdots$						
	$y_j$			$f(x, y)$			
	$\vdots$						
	$y_n$						
	Total of frequencies of $X$ $f(x)$	$f(x) = \sum_y f(x, y)$					$N \rightarrow \sum_x \sum_y f(x, y)$ $\sum_y \sum_x f(x, y)$

**Example 10.14.** The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test.

Marks ↓	Ages in years →	18	19	20	21	Total
		10–20	2	2	—	8
20–30	5	4	6	4	19	
30–40	6	8	10	11	35	
40–50	4	4	6	8	22	
50–60	—	2	4	4	10	
60–70	—	2	3	1	6	
Total	19	22	31	28	100	

Calculate the correlation coefficient.

**Solution.**

CORRELATION TABLE

v y	x Marks	u	-1	0	1	2	Total $f(v)$	$v f(v)$	$v^2 f(v)$	$\sum u^2 f(u, v)$	
			18	19	20	21					
-2 15	10—20	(8)	(0)	(-4)			8	-16	32	4	
		4	2	2							
-1 25	20—30	(5)	(0)	(-5)	(-8)		10	-19	19	-9	
		5	4	6	4						
0 35	30—40	(0)	(0)	(0)	(0)		35	0	0	0	
		6	8	10	11						
1 45	40—50	(-4)	(0)	(6)	(16)		22	22	22	18	
		4	4	6	8						
2 55	50—60		(0)	(8)	(16)		10	20	40	24	
			2	4	4						
3 65	60—70		(0)	(9)	(6)		6	18	54	15	
			2	3	1						
Total $f(u)$		19	22	31	28	100	25	167	-52		
$u f(u)$		-19	0	31	56	68					
$u^2 f(u)$		19	0	31	112	162					
$u \sum_v v f(u, v)$		9	0	13	30	52					

Let

$$U = X - 19, V = \{(Y - 35)/10\}$$

$$\bar{u} = \frac{1}{N} \sum_u u f(u) = \frac{68}{100} = 0.68, \bar{v} = \frac{1}{N} \sum_v v g(v) = \frac{25}{100} = 0.25$$

$$\text{Cov}(u, v) = \frac{1}{N} \sum_u \sum_v u v f(u, v) - \bar{u} \bar{v} = \frac{1}{100} \times 52 - 0.68 \times 0.25 = 0.35$$

$$\sigma_u^2 = \frac{1}{N} \sum_u u^2 f(u) - \bar{u}^2 = \frac{162}{100} - (0.68)^2 = 1.1576$$

$$\sigma_v^2 = \frac{1}{N} \sum_v v^2 g(v) - \bar{v}^2 = \frac{167}{100} - (0.25)^2 = 1.6075$$

$$\therefore r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_u \sigma_v} = \frac{0.35}{\sqrt{1.1576 \times 1.6075}} = 0.25$$

Since correlation coefficient is independent of change of origin and scale,  
 $r(X, Y) = r(U, V) = 0.25$

**Remark.** Figures in circles in the table on page 10.34 are the product terms  $uvf(u, v)$ :

**Example 10.15.** The joint probability distribution of  $X$  and  $Y$  is given below:

$X$	-1	+1	
$Y$			
0	$\frac{1}{8}$	$\frac{3}{8}$	
1	$\frac{2}{8}$	$\frac{2}{8}$	

Find the correlation coefficient between  $X$  and  $Y$ .

**Solution.**

#### COMPUTATION OF MARGINAL PROBABILITIES

$X$	-1	+1	$g(y)$
$Y$			
0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{4}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{5}{8}$
$p(x)$	$\frac{3}{8}$	$\frac{5}{8}$	1

We have :

$$E(X) = \sum xp(x) = (-1) \times \frac{3}{8} + 1 \times \frac{5}{8} = \frac{1}{4}$$

$$E(X^2) = \sum x^2 p(x) = (-1)^2 \times \frac{3}{8} + 1^2 \times \frac{5}{8} = 1$$

$$\therefore \text{Var}(X) = E(X^2) - [E(X)]^2 = 1 - \frac{1}{16} = \frac{15}{16}$$

$$E(Y) = \sum y g(y) = 0 \times \frac{4}{8} + 1 \times \frac{5}{8} = \frac{1}{2}$$

$$E(Y^2) = \sum y^2 g(y) = 0^2 \times \frac{4}{8} + 1^2 \times \frac{5}{8} = \frac{5}{8}$$

$$\therefore \text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{5}{8} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

$$\begin{aligned}
 E(XY) &= 0 \times (-1) \times \frac{1}{8} + 0 \times 1 \times \frac{3}{8} + 1 \times (-1) \times \frac{2}{8} + 1 \times 1 \times \frac{2}{8} \\
 &= -\frac{2}{8} + \frac{2}{8} = 0
 \end{aligned}$$

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0 - \frac{1}{4} \times \frac{1}{2} = -\frac{1}{8}$$

$$\therefore r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{-\frac{1}{8}}{\sqrt{\frac{15}{16} \times \frac{1}{4}}} = \frac{-1}{\sqrt{15}} = \frac{-1}{3.873} \\ = -0.2582$$

**EXERCISE 10(b)**

1. Write a brief note on the correlation table :

The following are the marks obtained by 24 students in a class test of Statistics and Mathematics :

Role No. of Students	1	2	3	4	5	6	7	8	9	10	11	12
Marks in Statistics	15	0	1	3	16	2	18	5	4	17	6	19
Marks in Mathematics	13	1	2	7	8	9	12	9	17	16	6	18
Role No. of Students	13	14	15	16	17	18	19	20	21	22	23	24
Marks in Statistics	14	9	8	13	10	13	11	11	12	18	9	7
Marks in Mathematics	11	3	5	4	10	11	14	7	18	15	15	3

Prepare a correlation table taking the magnitude of each class interval as four marks and the first class interval as "equal to 0 and less than 4". Calculate Karl Pearson's coefficient of correlation between the marks in Statistics and marks in Mathematics from the correlation table.

Ans. 0.5544..

2. An employment bureau asked applicants their weekly wages on jobs last held. The actual wages were obtained for 54 of them; and are recorded in the table below;  $x$  represents reported wage,  $y$  actual wage, and the entry in the table represents frequency. Find the correlation coefficient and comment on the significance of the computed value. [Four figure log table may be used].

$y$ → $\downarrow x$	15	20	25	30	35	40
40						2
35				3	5	
30			4	15		
25				20		
20		3	1			
15	1					

[Calcutta Univ. B.Sc. (Maths. Hons.), 1986]

3. Calculate the correlation coefficient from the following table :—

$x$ → $\downarrow y$	0—10	10—20	20—30	30—40
0—5	1	3	2	0
5—10	7	10	8	1
10—15	10	13	10	8
15—20	5	8	10	7
20—25	0	1	5	4

4. (a) Find the correlation coefficient between age and salary of 50 workers in a factory :

Age (in years) ↓	Daily pay in rupees				
	160—169	170—179	180—189	190—199	200—209
20—30	5	3	1	...	...
30—40	2	6	2	1	...
40—50	1	2	4	2	2
50—60	...	1	3	6	2
60—70	...	...	1	1	5

(b) Find the coefficient of correlation between the ages of 100 mothers and daughters :

Age of mothers in years (X)	Age of daughters in years (Y)					Total
	5—10	10—15	15—20	20—25	25—30	
15—25	6	3				9
25—35	3	16	10			29
35—45		10	15	7		32
45—55			7	10	4	21
55—65				4	5	9
Total	9	29	32	21	9	100

[Madras Univ. B.Sc. (Main Maths.), 1991]

5. Given the following frequency distribution of (X, Y) :

X ↓	5	10	Total
Y ↓			
10	30	20	50
20	20	30	50
Total	50	50	100

Find the frequency distribution of (U, V), where

$$U = \frac{X - 7.5}{2.5}, V = \frac{Y - 15}{-5}.$$

What shall be the relationship between the correlation coefficients between X, Y, and U, V?

6. (a) Find the coefficient of correlation between X and Y for the following table :

		$y_1$	$y_2$	Total
		$p_{11}$	$p_{12}$	$P$
		$p_{21}$	$p_{22}$	$Q$
Total		$P'$	$Q'$	1

(b) Consider the following probability distribution :

		0	1	2
		0.1	0.2	0.1
		0.2	0.3	0.1

Calculate  $E(X)$ ,  $\text{Var}(X)$ ,  $\text{Cov}(X, Y)$  and  $r(X, Y)$ .

[Delhi Univ. M.A. (Eco.), 1991]

(c) Let  $(X, Y)$  have the p.m.f.

$$p(0, 1) = p(1, 0) = \frac{1}{3}; \quad p(0, -1) = p(-1, 0) = \frac{1}{6}.$$

Find  $r(X, Y)$ . Are  $X$  and  $Y$  independent ? For what values of  $k$ ,  $X + kY$  and  $kX + Y$  are uncorrelated ?

**10.5. Probable Error of Correlation Coefficient.** If  $r$  is the correlation coefficient in a sample of  $n$  pairs of observations, then its *standard error* is given by

$$\text{S.E.(}r\text{)} = \frac{1 - r^2}{\sqrt{n}}$$

Probable error of correlation coefficient is given by

$$\text{P.E.(}r\text{)} = 0.6745 \times \text{S.E.} = 0.6745 \frac{(1 - r^2)}{\sqrt{n}} \quad \dots(10.6)$$

Probable error is an old measure for testing the reliability of an observed correlation coefficient. The reason for taking the factor 0.6745 is that in a normal distribution, the range  $\mu \pm 0.6745 \sigma$  covers 50% of the total area. According to Secrist, "The probable error of the correlation co-efficient is an amount which if added to and subtracted from the mean correlation coefficient, produces amounts within which the chances are even that a coefficient of correlation from a series selected at random will fall."

If  $r < \text{P.E.(}r\text{)}$ , correlation is not at all significant. If  $r > 6 \text{ P.E.(}r\text{)}$ , it is definitely significant. A rigorous method (*t*-test) of testing the significance of an observed correlation coefficient will be discussed later in "tests of significance" in sampling [c.f. § 14.4-11].

Probable error also enables us to find the limits within which the population correlation coefficient can be expected to vary. The limits are  $r \pm P.E.(r)$ .

**10.6. Rank Correlation.** Let us suppose that a group of  $n$  individuals is arranged in order of merit or proficiency in possession of two characteristics  $A$  and  $B$ . These ranks in the two characteristics will, in general, be different. For example, if we consider the relation between intelligence and beauty, it is not necessary that a beautiful individual is intelligent also. Let  $(x_i, y_i); i = 1, 2, \dots, n$  be the ranks of the  $i$ th individual in two characteristics  $A$  and  $B$  respectively. Pearsonian coefficient of correlation between the ranks  $x_i$ 's and  $y_i$ 's is called the rank correlation coefficient between  $A$  and  $B$  for that group of individuals.

Assuming that no two individuals are bracketed equal in either classification, each of the variables  $X$  and  $Y$  takes the values  $1, 2, \dots, n$ .

$$\text{Hence } \bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + 3 + \dots + n) = \frac{n+1}{2}$$

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

$$= \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

$$\therefore \sigma_x^2 = \frac{n^2-1}{12} = \sigma_y^2$$

In general  $x_i \neq y_i$ . Let  $d_i = x_i - y_i$

$$\therefore d_i = (x_i - \bar{x}) - (y_i - \bar{y}) \quad (\because \bar{x} = \bar{y})$$

Squaring and summing over  $i$  from 1 to  $n$ , we get

$$\begin{aligned} \sum d_i^2 &= \sum ((x_i - \bar{x}) - (y_i - \bar{y}))^2 \\ &= \sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2 - 2\sum (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

Dividing both sides by  $n$ , we get

$$\frac{1}{n} \sum d_i^2 = \sigma_x^2 + \sigma_y^2 - 2 \operatorname{Cov}(X, Y) = \sigma_x^2 + \sigma_y^2 - 2\rho \sigma_x \sigma_y,$$

where  $\rho$  is the rank correlation coefficient between  $A$  and  $B$ .

$$\begin{aligned} \therefore \frac{1}{n} \sum d_i^2 &= 2\sigma_x^2 - 2\rho \sigma_x^2 \Rightarrow 1 - \rho = \frac{\sum d_i^2}{2n\sigma_x^2} \\ \Rightarrow \rho &= 1 - \frac{\sum d_i^2}{2n\sigma_x^2} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \quad \dots(10.7) \end{aligned}$$

which is the *Spearman's formula for the rank correlation coefficient*.

**Remark.** We always have

$$\sum d_i = \sum (x_i - y_i) = \sum x_i - \sum y_i = n(\bar{x} - \bar{y}) = 0 \quad (\because \bar{x} = \bar{y})$$

This serves as a check on the calculations.

**10-6-1. Tied Ranks.** If some of the individuals receive the same rank in a ranking of merit, they are said to be tied. Let us suppose that  $m$  of the individuals, say,  $(k+1)^{th}$ ,  $(k+2)^{th}$ , ...,  $(k+m)^{th}$  are tied. Then each of these  $m$  individuals is assigned a common rank, which is the arithmetic mean of the ranks  $k+1, k+2, \dots, k+m$ .

*Derivation of  $\rho(X, Y)$ :* We have :

$$\rho(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{[\sum (X - \bar{X})^2 \cdot \sum (Y - \bar{Y})^2]^{1/2}} = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}} \quad \dots (*)$$

where  $x = X - \bar{X}$ ,  $y = Y - \bar{Y}$ .

If  $X$  and  $Y$  each takes the values  $1, 2, \dots, n$ , then we have

$$\bar{X} = (n+1)/2 = \bar{Y}$$

$$\text{and } n\sigma_X^2 = \sum x^2 = \frac{n(n^2-1)}{12} \text{ and } n\sigma_Y^2 = \sum y^2 = \frac{n(n^2-1)}{12} \quad \dots (**)$$

$$\text{Also } \sum d^2 = \sum (X - Y)^2 = \sum [(X - \bar{X}) - (Y - \bar{Y})]^2 = \sum (x - y)^2$$

$$\Rightarrow \sum d^2 = \sum x^2 + \sum y^2 - 2\sum xy$$

$$\Rightarrow \sum xy = \frac{1}{2} [\sum x^2 + \sum y^2 - \sum d^2] \quad \dots (***)$$

We shall now investigate the effect of common ranking, (in case of ties), on the sum of squares of the ranks. Let  $S^2$  and  $S_1^2$  denote the sum of the squares of untied and tied ranks respectively.

Then we have :

$$\begin{aligned} S^2 &= (k+1)^2 + (k+2)^2 + \dots + (k+m)^2 \\ &= mk^2 + (1^2 + 2^2 + \dots + m^2) + 2k(1+2+\dots+m) \\ &= mk^2 + \frac{m(m+1)(2m+1)}{6} + mk(m+1) \end{aligned}$$

$$\begin{aligned} S_1^2 &= m(\text{Average rank})^2 \\ &= m \left[ \frac{(k+1) + (k+2) + \dots + (k+m)}{m} \right]^2 \\ &= m \left( k + \frac{m+1}{2} \right)^2 = mk^2 + \frac{m(m+1)^2}{4} + m k (m+1) \end{aligned}$$

$$\therefore S^2 - S_1^2 = \frac{m(m+1)}{12} [2(2m+1) - 3(m+1)] = \frac{m(m^2-1)}{12}$$

Thus the effect of tying  $m$  individuals (ranks) is to reduce the sum of the squares by  $m(m^2-1)/12$ , though the mean value of the ranks remains the same, viz.,  $(n+1)/2$ .

Suppose that there are  $s$  such sets of ranks to be tied in the  $X$ -series so that the total sum of squares due to them is

$$\frac{1}{12} \sum_{i=1}^s m_i (m_i^2 - 1) = \frac{1}{12} \sum_{i=1}^s (m_i^3 - m_i) = T_X, (\text{say}) \quad \dots (10.7a)$$

Similarly suppose that there are  $t$  such sets of ranks to be tied with respect to the other series  $Y$  so that sum of squares due to them is :

$$\frac{1}{12} \sum_{j=1}^t m_j' \cdot (m_j'^2 - 1) = \frac{1}{12} \sum_{j=1}^t (m_j'^3 - m_j') = T_Y, \text{ (say)} \quad \dots(10.7b)$$

Thus, in the case of ties, the new sums of squares are given by :

$$n \operatorname{Var}'(X) = \sum x^2 - T_X = \frac{n(n^2 - 1)}{12} - T_X$$

$$n \operatorname{Var}'(Y) = \sum y^2 - T_Y = \frac{n(n^2 - 1)}{12} - T_Y$$

$$\text{and } n \operatorname{Cov}'(X, Y) = \frac{1}{2} [\sum x^2 - T_X + \sum y^2 - T_Y - \sum d^2] \quad [\text{From } (***)]$$

$$= \frac{1}{2} \left[ \frac{n(n^2 - 1)}{12} - T_X + \frac{n(n^2 - 1)}{12} - T_Y - \sum d^2 \right]$$

$$= \frac{n(n^2 - 1)}{12} - \frac{1}{2} [(T_X + T_Y) + \sum d^2]$$

$$\rho(X, Y) = \frac{\frac{n(n^2 - 1)}{12} - \frac{1}{2} [T_X + T_Y + \sum d^2]}{\left[ \frac{n(n^2 - 1)}{12} - T_X \right]^{1/2} \left[ \frac{n(n^2 - 1)}{12} - T_Y \right]^{1/2}}$$

$$= \frac{\frac{n(n^2 - 1)}{6} - [\sum d^2 + T_X + T_Y]}{\left[ \frac{n(n^2 - 1)}{6} - 2T_X \right]^{1/2} \left[ \frac{n(n^2 - 1)}{6} - 2T_Y \right]^{1/2}}$$

...(10.7c)

where  $T_X$  and  $T_Y$  are given by (10.7a) and (10.7b).

**Remark.** If we adjust only the covariance term i.e.,  $\sum xy$  and not the variances  $\sigma_X^2$  (or  $\sum x^2$ ) and  $\sigma_Y^2$  (or  $\sum y^2$ ) for ties, then the formula (10.7c) reduces to :

$$\begin{aligned} \rho(X, Y) &= \frac{\frac{n(n^2 - 1)}{6} - (\sum d^2 + T_X + T_Y)}{n(n^2 - 1)/6} \\ &= 1 - \frac{6 [\sum d^2 + T_X + T_Y]}{n(n^2 - 1)}, \end{aligned} \quad \dots(10.7d)$$

a formula which is commonly used in practice for numerical problems. For illustration, see Example 10.18.

**Example 10.16.** The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics.

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

Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Physics.

**Solution.**

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

Rank correlation coefficient is given by

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 136}{16 \times 255} = 1 - \frac{1}{5} = \frac{4}{5} = 0.8$$

**Example 10-17.** Ten competitors in a musical test were ranked by the three judges A, B and C in the following order :

Ranks by A : 1 6 5 10 3 2 4 9 7 8

Ranks by B : 3 5 8 4 7 10 2 1 6 9

Ranks by C : 6 4 9 8 1 2 3 10 5 7

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

**Solution.** Here  $n = 10$

Ranks by A (X)	Ranks by B (Y)	Ranks by C (Z)	$d_1 = X - Y$	$d_2 = X - Z$	$d_3 = Y - Z$	$d_1^2$	$d_2^2$	$d_3^2$
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	-2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
Total			$\sum d_1 = 0$	$\sum d_2 = 0$	$\sum d_3 = 0$	$\sum d_1^2 = 200$	$\sum d_2^2 = 60$	$\sum d_3^2 = 214$

$$\rho(X, Y) = 1 - \frac{6 \sum d_1^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = -\frac{7}{33}$$

$$\rho(X, Z) = 1 - \frac{6 \sum d_2^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$\rho(Y, Z) = 1 - \frac{6 \sum d_3^2}{n(n^2 - 1)} = 1 - \frac{6 \times 214}{10 \times 99} = -\frac{49}{165}$$

Since  $\rho(X, Z)$  is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

**10-6-2. Repeated Ranks (Continued).** If any two or more individuals are bracketed equal in any classification with respect to characteristics A and B, or if there is more than one item with the same value in the series, then the Spearman's formula (10-7) for calculating the rank correlation coefficient breaks down, since in this case each of the variables X and Y does not assume the values 1, 2, ..., n and consequently,  $\bar{x} \neq \bar{y}$ .

In this case, common ranks are given to the repeated items. This common rank is the average of the ranks which these items would have assumed if they were slightly different from each other and the next item will get the rank next to the ranks already assumed. As a result of this, following adjustment or correction is made in the rank correlation formula [c.f. (10-7c) and (10-7d)].

In the formula, we add the factor  $\frac{m(m^2 - 1)}{12}$  to  $\sum d^2$ , where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the X-series and Y-series.

**Example 10-18.** Obtain the rank correlation coefficient for the following data:

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

**Solution.**

#### CALCULATIONS FOR RANK CORRELATION

X	Y	Rank X (x)	Rank Y (y)	d = x - y	d <sup>2</sup>
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
$\Sigma d = 0$					$\Sigma d^2 = 72$

In the X-series we see that the value 75 occurs 2 times. The common rank given to these values is 2.5 which is the average of 2 and 3, the ranks which these values would have taken if they were different. The next value 68, then gets the next rank which is 4. Again we see that value 64 occurs thrice. The common rank given to it is 6 which is the average of 5, 6 and 7. Similarly in

the  $Y$ -series, the value 68 occurs twice and its common rank is 3.5 which is the average of 3 and 4. As a result of these common rankings, the formula for ' $\rho$ ' has to be corrected. To  $\sum d^2$  we add  $\frac{m(m^2 - 1)}{12}$  for each value repeated, where

$m$  is the number of times a value occurs. In the  $X$ -series the correction is to be applied twice, once for the value 75 which occurs twice ( $m = 2$ ) and then for the value 64 which occurs thrice ( $m = 3$ ). The total correction for the  $X$ -series is

$$\frac{2(4 - 1)}{12} + \frac{3(9 - 1)}{12} = \frac{5}{2}$$

Similarly, this correction for the  $Y$ -series is  $\frac{2(4 - 1)}{12} = \frac{1}{2}$ , as the value 68 occurs twice.

$$\text{Thus } \rho = 1 - \frac{6 \left[ \sum d^2 + \frac{5}{2} + \frac{1}{2} \right]}{n(n^2 - 1)} = 1 - \frac{6(72 + 3)}{10 \times 99} = 0.545$$

**10.6.3. Limits for the Rank Correlation Coefficient.** Spearman's rank correlation coefficient is given by

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

' $\rho$ ' is maximum, if  $\sum_{i=1}^n d_i^2$  is minimum, i.e., if each of the deviations  $d_i$  is minimum. But the minimum value of  $d_i$  is zero in the particular case  $x_i = y_i$ , i.e., if the ranks of the  $i$ th individual in the two characteristic are equal. Hence the maximum value of  $\rho$  is +1, i.e.,  $\rho \leq 1$ .

' $\rho$ ' is minimum, if  $\sum_{i=1}^n d_i^2$  is maximum, i.e., if each of the deviations  $d_i$  is maximum which is so if the ranks of the  $n$  individuals in the two characteristics are in the opposite directions as given below :

$x$	1	2	3	...	...	$n - 1$	$n$	
$y$	$n$	$n - 1$	$n - 2$	...	...	2	1	...(*)

**Case 1.** Suppose  $n$  is odd and equal to  $(2m + 1)$  then the values of  $d$  are :

$d : 2m, 2m - 2, 2m - 4, \dots, 2, 0, -2, -4, \dots, -(2m - 2), -2m$ .

$$\therefore \sum_{i=1}^n d_i^2 = 2 \{ (2m)^2 + (2m - 2)^2 + \dots + 4^2 + 2^2 \}$$

$$= 8 \{ m^2 + (m - 1)^2 + \dots + 1^2 \} = \frac{8m(m + 1)(2m + 1)}{6}$$

$$\text{Hence } \rho = 1 - \frac{\sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{8m(m+1)(2m+1)}{(2m+1)\{(2m+1)^2 - 1\}} \\ = \frac{8m(m+1)}{(4m^2 + 4m)} = 1 - \frac{8m(m+1)}{4m(m+1)} = -1$$

**Case II.** Let  $n$  be even and equal to  $2m$ , (say).

Then the values of  $d$  are

$$(2m-1), (2m-3), \dots, 1, -1, -3, \dots, -(2m-3), -(2m-1) \\ \therefore \sum d_i^2 = 2\{(2m-1)^2 + (2m-3)^2 + \dots + 1^2\} \\ = 2[(\{(2m)^2 + (2m-1)^2 + (2m-2)^2 + \dots + 2^2 + 1^2\}) \\ - \{(2m)^2 + (2m-2)^2 + \dots + 4^2 + 2^2\}] \\ = 2[1^2 + 2^2 + \dots + (2m)^2 - \{2^2m^2 + 2^2(m-1)^2 + \dots + 2^2\}] \\ = 2\left[\frac{2m(2m+1)(4m+1)}{6} - 4\frac{m(m+1)(2m+1)}{6}\right] \\ = \frac{2m}{3}[(2m+1)(4m+1) - 2(m+1)(2m+1)] \\ = \frac{2m}{3}[(2m+1)(4m+1) - 2m(m+1)] \\ = \frac{2m}{3}(2m+1)(2m-1) = \frac{2m(4m^2-1)}{3} \\ \therefore \rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} = 1 - \frac{4m(4m^2-1)}{2m(4m^2-1)} = -1$$

Thus the limits for rank correlation coefficient are given by  $-1 \leq \rho \leq 1$ .

**Aliter.** For an alternate and simpler proof for obtaining the minimum value of  $\rho$ , from (\*) onward, proceed as in Hint to Question Number 9 of Exercise 10(c).

#### Remarks on Spearman's Rank Correlation Coefficient.

1.  $\sum d = \sum x - \sum y = n(\bar{x} - \bar{y}) = 0$ , which provides a check for numerical calculations.

2. Since Spearman's rank correlation coefficient  $\rho$  is nothing but Pearsonian correlation coefficient between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.

3. Karl Pearson's correlation coefficient assume that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure which is distribution free (or non-parametric). A distribution-free measure is one which does not make any assumptions about the parameters of the population. Spearman's  $\rho$  is such a measure (*i.e.*, distribution-free), since no strict assumptions are made about the form of the population from which sample observations are drawn.

4. Spearman's formula is easy to understand and apply as compared with Karl Pearson's formula. The value obtained by the two formulae, *viz.*, Pearsonian  $r$  and Spearman's  $\rho$ , are generally different. The difference arises due to the fact that when ranking is used instead of full set of observations, there is

always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.

6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution (Correlation Table). For  $n > 30$ , this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time-consuming.

### EXERCISE 10(c)

1. Prove that Spearman's rank correlation coefficient is given by

$$1 - \frac{6 \sum d_i^2}{n^3 - n}$$
, where  $d_i$  denotes the difference between the ranks of  $i$ th individual.

2. (a) Explain the difference between product moment correlation coefficient and rank correlation coefficient.

- (b) The rankings of ten students in two subjects  $A$  and  $B$  are as follows :

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

Find the correlation coefficient.

3. (a) Calculate the coefficient of correlation for ranks from the following data :

$$(X, Y) : (5, 8), (10, 3), (6, 2), (3, 9), (19, 12), (5, 3), (6, 17), (12, 18), (8, 22), (2, 12), (10, 17), (19, 20).$$

[Calicut Univ. B.Sc. (Subs. Stat.), Oct. 1991]

- (b) Ten recruits were subjected to a selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test.

The marks secured by recruits in the selection test ( $X$ ) and in the proficiency test ( $Y$ ) are given below :—

Serial No. :	1	2	3	4	5	6	7	8	9	10
$X$ :	10	15	12	17	13	16	24	14	22	20
$Y$ :	30	42	45	46	33	34	40	35	39	38

Calculate product moment correlation coefficient and rank correlation coefficient. Why are two coefficients different ?

4. (a) The I.Q.'s of a group of 6 persons were measured, and they then sat for a certain examination. Their I.Q.'s and examination marks were as follows :

Person :	$A$	$B$	$C$	$D$	$E$	$F$
I.Q. :	110	100	140	120	80	90
Exam. marks :	70	60	80	60	10	20

Compute the coefficients of correlation and rank correlation. Why are the correlation figures obtained different ?

Ans. 0.882 and 0.9.

The difference arises due to the fact that when ranking is used instead of the full set of observations, there is always some loss of information.

(b) The value of ordinary correlation ( $r$ ) for the following data is 0.636 :—

$X$  : .05 .14 .24 .30 .47 .52 .57 .61 .67 .72

$Y$  : 1.08 1.15 1.27 1.33 1.41 1.46 1.54 2.72 4.01 9.63

(i) Calculate Spearman's rank-correlation ( $\rho$ ) for this data.

(ii) What advantage of  $\rho$  was brought out in this example ?

4. Ten competitors in a beauty contest are ranked by three judges as follows :

	<i>Competitors</i>									
<i>Judges</i>	1	2	3	4	5	6	7	8	9	10
<i>A</i>	6	5	3	10	2	4	9	7	8	1
<i>B</i>	5	8	4	7	10	2	1	6	9	3
<i>C</i>	4	9	8	1	2	3	10	5	7	6

Discuss which pair of judges has the nearest approach to common tastes of beauty.

5. A sample of 12 fathers and their eldest sons gave the following data about their height in inches :

Father : 65 63 67 64 68 62 70 66 68 67 69 71

Son : 68 66 68 65 69 66 68 65 71 67 68 70

Calculate coefficient of rank correlation. (Ans. 0.7220)

6. The coefficient of rank correlation between marks in Statistics and marks in Mathematics obtained by a certain group of students is 0.8. If the sum of the squares of the difference in ranks is given to be 33, find the number of student in the group (Ans. 10). [Madras Univ. B.Sc., 1990]

7. The coefficient of rank correlation of the marks obtained by 10 students in Maths and Statistics was found to be 0.5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

$$\text{Hint.} \quad 0.5 = 1 - \frac{6 \sum d^2}{10 \times 99}$$

$$\Rightarrow \quad \sum d^2 = \frac{990}{6 \times 2} = 82.5$$

Since one difference was wrongly taken as 3 instead of 7, the correct value of  $\sum d^2$  is given by

$$\text{Corrected } \sum d^2 = 82.5 - (3)^2 + (7)^2 = 122.5$$

$$\therefore \quad \text{Corrected } \rho = 1 - \frac{6 \times 122.5}{10 \times 99} = 0.2576$$

8. If  $d_i$  be the difference in the ranks of the  $i$ th individual in two different characteristics, then show that the maximum value of  $\sum_{i=1}^n d_i^2$  is  $\frac{1}{3}(n^3 - n)$ .

Hence or otherwise, show that rank correlation coefficient lies between -1 and +1.  
[Delhi Univ. B.Sc. (Maths. Hons.), 1986]

9. Let  $x_1, x_2, \dots, x_n$  be the ranks of  $n$  individuals according to a character  $A$  and  $y_1, y_2, \dots, y_n$  be the ranks of the same individuals according to other character  $B$ . Obviously  $(x_1, x_2, \dots, x_n)$  and  $(y_1, y_2, \dots, y_n)$  are permutations of  $1, 2, \dots, n$ . It is given that  $x_i + y_i = 1 + n$ , for  $i = 1, 2, \dots, n$ . Show that the value of the rank correlation coefficient  $\rho$  between the characters  $A$  and  $B$  is -1.

**Hint.** We are given  $x_i + y_i = n + 1 \forall i = 1, 2, \dots, n$

Also  $x_i - y_i = d_i$

$$\therefore 2x_i = n + 1 + d_i \Rightarrow d_i = 2x_i - (n + 1)$$

$$\therefore \sum_{i=1}^n d_i^2 = \sum_{i=1}^n [4x_i^2 + (n+1)^2 - 2(n+1)2x_i]$$

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

$$= \frac{n(n^2-1)}{3}$$

$$\therefore \rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} = -1$$

**Remark.** From Spearman's formula we note that  $\rho$  will be minimum if  $\sum d_i^2$  is maximum, which will be so if the ranks  $X$  and  $Y$  are in opposite directions as given below :

$x$	1	2	3	$\dots$	$n$
$y$	$n$	$n-1$	$n-2$	$\dots$	1

This gives us

$$x_i + y_i = n + 1, i = 1, 2, \dots, n.$$

Hence the value of  $\rho = -1$  obtained above is minimum value of  $\rho$ .

10. Show that in a ranked bivariate distribution in which no ties occur and in which the variables are independent

(a)  $\sum_i d_i^2$  is always even, and

(b) there are not more than  $\frac{1}{6}(n^3 - n) + 1$  possible values of  $r$ .

11. Show that if  $X, Y$  be identically distributed with common probability mass function :  $P(X = k) = \frac{1}{N}$ , for  $k = 1, 2, \dots, N; N > 1$ ,

then  $\rho_{X,Y}$ , the correlation coefficient between  $X$  and  $Y$ , is given by

$$1 - \frac{6E(X - Y)^2}{N^2 - 1}$$

[Delhi Univ. B.Sc. (Maths Hons.), 1992]

**10.7. Regression.** The term "regression" literally means "stepping back towards the average". It was first used by a British biometrician Sir Francis Galton (1822—1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to "regress" or "step back" to the average population height. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry.

**Definition.** *Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.*

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable* and the variable which influences the values or is used for prediction, is called *independent variable*. In regression analysis independent variable is also known as *regressor or predictor or explanatory variable* while the dependent variable is also known as *regressed or explained variable*.

**10.7.1. Lines of Regression.** If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round some curve called the "curve of regression". If the curve is a straight line, it is called the line of regression and there is said to be *linear regression* between the variables, otherwise regression is said to be *curvilinear*.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of "best fit" and is obtained by the *principles of least squares*.

Let us suppose that in the bivariate distribution  $(x_i, y_i); i = 1, 2, \dots, n$ ;  $Y$  is dependent variable and  $X$  is independent variable. Let the line of regression of  $Y$  on  $X$  be  $Y = a + bX$ .

According to the principle of least squares, the normal equations for estimating  $a$  and  $b$  are (c.f. (9.2a))

$$\sum_{i=1}^n y_i = na + b \sum_{i=1}^n x_i \quad \dots(10-8)$$

and  $\sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 \quad \dots(10-9)$

From (10-8) on dividing by  $n$ , we get

$$\bar{y} = a + b\bar{x} \quad \dots(10-10)$$

Thus the line of regression of  $Y$  on  $X$  passes through the point  $(\bar{x}, \bar{y})$ .

Now

$$\mu_{11} = \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \Rightarrow \frac{1}{n} \sum_{i=1}^n x_i y_i = \mu_{11} + \bar{x} \bar{y} \quad \dots(10-11)$$

$$\text{Also } \sigma_X^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \Rightarrow \frac{1}{n} \sum_{i=1}^n x_i^2 = \sigma_X^2 + \bar{x}^2 \quad \dots(10-11a)$$

Dividing (10-9) by  $n$  and using (10-11) and (10-11a), we get

$$\mu_{11} + \bar{x} \bar{y} = a\bar{x} + b(\sigma_X^2 + \bar{x}^2) \quad \dots(10-12)$$

Multiplying (10-10) by  $\bar{x}$  and then subtracting from (10-12), we get

$$\mu_{11} = b\sigma_X^2 \Rightarrow b = \frac{\mu_{11}}{\sigma_X^2} \quad \dots(10-13)$$

Since ' $b$ ' is the slope of the line of regression of  $Y$  on  $X$  and since the line of regression passes through the point  $(\bar{x}, \bar{y})$ , its equation is

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2} (X - \bar{x}) \quad \dots(10-14)$$

$$\Rightarrow Y - \bar{y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{x}) \quad \dots(10-14a)$$

Starting with the equation  $X = A + BY$  and proceeding similarly or by simply interchanging the variables  $X$  and  $Y$  in (10-14) and (10-14a), the equation of the line of regression of  $X$  on  $Y$  becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2} (Y - \bar{y}) \quad \dots(10-15)$$

$$\Rightarrow X - \bar{x} = r \frac{\sigma_X}{\sigma_Y} (Y - \bar{y}) \quad \dots(10-15a)$$

Aliter. The straight line defined by

$$Y = a + bX \quad \dots(i)$$

and satisfying the residual (least square) condition

$$S = E [(Y - a - bX)^2] = \text{Minimum} \quad \dots(10-16)$$

for variations in  $a$  and  $b$ , is called the line of regression of  $Y$  on  $X$ .

The necessary and sufficient conditions for a minima of  $S$ , subject to variations in  $a$  and  $b$  are :

$$(i) \frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0 \quad \text{and} \quad \dots (*)$$

$$(ii) \Delta = \begin{vmatrix} \frac{\partial^2 S}{\partial a^2} & \frac{\partial^2 S}{\partial a \partial b} \\ \frac{\partial^2 S}{\partial b \partial a} & \frac{\partial^2 S}{\partial b^2} \end{vmatrix} > 0 \quad \text{and} \quad \frac{\partial^2 S}{\partial a^2} > 0 \quad \dots (**)$$

Using (\*), we get

$$\frac{\partial S}{\partial a} = -2 E [Y - a - bX] = 0 \quad \dots (iii)$$

$$\frac{\partial S}{\partial b} = -2 E [X(Y - a - bX)] = 0 \quad \dots (iv)$$

$$\Rightarrow E(Y) = a + bE(X) \dots (v) \quad \text{and} \quad E(XY) = aE(X) + bE(X^2) \quad \dots (vi)$$

Equation (v) implies that the line (i) of regression of  $Y$  on  $X$  passes through the mean value  $[E(X), E(Y)]$ .

Multiplying (v) by  $E(X)$  and subtracting from (vi), we get

$$E(XY) - E(X)E(Y) = b[E(X^2) - \{E(X)\}^2]$$

$$\Rightarrow \text{Cov}(X, Y) = b \sigma_X^2 \Rightarrow b = \frac{\text{Cov}(X, Y)}{\sigma_X^2} = \frac{r \sigma_Y}{\sigma_X} \quad \dots (vii)$$

Subtracting (v) from (i) and using (vii), we obtain the equation of line of regression of  $Y$  on  $X$  as :

$$Y - E(Y) = \frac{\text{Cov}(X, Y)}{\sigma_X^2} [X - E(X)] \Rightarrow Y - E(Y) = \frac{r \sigma_Y}{\sigma_X} [X - E(X)]$$

Similarly, the straight line defined by  $X = A + BY$   
and satisfying the residual condition

$$E[X - A - BY]^2 = \text{Minimum},$$

is called the line of regression of  $X$  on  $Y$ .

**Remarks 1.** We note that

$$\frac{\partial^2 S}{\partial a^2} = 2 > 0, \text{ and}$$

$$\frac{\partial^2 S}{\partial b^2} = 2E(X^2) \quad \text{and} \quad \frac{\partial^2 S}{\partial a \partial b} = 2E(X)$$

Substituting in (\*\*), we have

$$\begin{aligned} \Delta &= \frac{\partial^2 S}{\partial a^2} \cdot \frac{\partial^2 S}{\partial b^2} - \left( \frac{\partial^2 S}{\partial a \partial b} \right)^2 \\ &= 4 [E(X^2) - \{E(X)\}^2] = 4 \cdot \sigma_X^2 > 0 \end{aligned}$$

Hence the solution of the least square equations (iii) and (iv), in fact, provides a minima of  $S$ .

2. The regression equation (10-14a) implies that the line of regression of  $Y$  on  $X$  passes through the point  $(\bar{x}, \bar{y})$ . Similarly (10-15a) implies that the line of regression of  $X$  on  $Y$  also passes through the point  $(\bar{x}, \bar{y})$ . Hence both the lines of regression pass through the point  $(\bar{x}, \bar{y})$ . In other words, the mean

values ( $\bar{x}, \bar{y}$ ) can be obtained as the point of intersection of the two regression lines.

**3. Why two lines of Regression?** There are always two lines of regression, one of  $Y$  on  $X$  and the other of  $X$  on  $Y$ . The line of regression of  $Y$  on  $X$  (10.14a) is used to estimate or predict the value of  $Y$  for any given value of  $X$ , i.e., when  $Y$  is a dependent variable and  $X$  is an independent variable. The estimate so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of least squares. We can also obtain an estimate of  $X$  for any given value of  $Y$  by using equation (10.14a) but the estimate so obtained will not be best since (10.14a) is obtained on minimising the sum of the squares of errors of estimates in  $Y$  and not in  $X$ . Hence to estimate or predict  $X$  for any given value of  $Y$ , we use the regression equation of  $X$  on  $Y$  (10.15a) which is derived on minimising the sum of the squares of errors of estimates in  $X$ . Here  $X$  is a dependent variable and  $Y$  is an independent variable. The two regression equations are not reversible or interchangeable because of the simple reason that the basis and assumptions for deriving these equations are quite different. The regression equation of  $Y$  on  $X$  is obtained on minimising the sum of the squares of the errors parallel to the  $Y$ -axis while the regression equation of  $X$  on  $Y$  is obtained on minimising the sum of squares of the errors parallel to the  $X$ -axis.

In a particular case of perfect correlation, positive or negative, i.e.,  $r \pm 1$ , the equation of line of regression of  $Y$  on  $X$  becomes :

$$\begin{aligned} Y - \bar{y} &= \pm \frac{\sigma_x}{\sigma_y} (X - \bar{x}) \\ \Rightarrow \quad \frac{Y - \bar{y}}{\sigma_y} &= \pm \left( \frac{X - \bar{x}}{\sigma_x} \right) \end{aligned} \quad \dots(10.16)$$

Similarly, the equation of the line of regression of  $X$  on  $Y$  becomes :

$$\begin{aligned} X - \bar{x} &= \pm \frac{\sigma_x}{\sigma_y} (Y - \bar{y}) \\ \Rightarrow \quad \frac{X - \bar{x}}{\sigma_x} &= \pm \left( \frac{Y - \bar{y}}{\sigma_y} \right), \end{aligned}$$

which is same as (10.16).

Hence in case of perfect correlation, ( $r = \pm 1$ ), both the lines of regression coincide. Therefore, in general, we always have two lines of regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.

**10.7.2. Regression Curves.** In modern terminology, the conditional mean  $E(Y | X = x)$  for a continuous distribution is called the regression function of  $Y$  on  $X$  and the graph of this function of  $x$  is known as the regression curve of  $Y$  on  $X$  or sometimes the regression curve for the mean of  $Y$ . Geometrically, the regression function represents the  $y$  co-ordinate of the centre of mass of the bivariate probability mass in the infinitesimal vertical strip bounded by  $x$  and  $x + dx$ .

Similarly, the regression function of  $X$  on  $Y$  is  $E(X|Y=y)$  and the graph of this function of  $y$  is called the regression curve (of the mean) of  $X$  on  $Y$ .

In case a regression curve is a straight line, the corresponding regression is said to be *linear*. If one of the regressions is linear, it does not however follow that the other is also linear. For illustration, See Example 10.21.

**Theorem 10.4.** Let  $(X, Y)$  be a two-dimensional random variable with  $E(X) = \bar{X}$ ,  $E(Y) = \bar{Y}$ ,  $V(X) = \sigma_x^2$ ,  $V(Y) = \sigma_y^2$  and let  $r = r(X, Y)$  be the correlation coefficient between  $X$  and  $Y$ . If the regression of  $Y$  on  $X$  is linear then

$$E(Y|X) = \bar{Y} + r \frac{\sigma_y}{\sigma_x} (X - \bar{X}) \quad \dots(10.16a)$$

Similarly, if the regression of  $X$  on  $Y$  is linear, then

$$E(X|Y) = \bar{X} + r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y}) \quad \dots(10.16b)$$

**Proof.** Let the regression equation of  $Y$  on  $X$  be

$$E(Y|x) = a + bx \quad \dots(1)$$

But by definition,

$$\begin{aligned} E(Y|x) &= \int_{-\infty}^{\infty} y f(y|x) dy = \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_X(x)} dy \\ \therefore \quad \frac{1}{f_X(x)} \int_{-\infty}^{\infty} y f(x,y) dy &= a + bx \end{aligned} \quad \dots(2)$$

Multiplying both sides of (2) by  $f_X(x)$  and integrating w.r.t.  $x$ , we get

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y) dy dx &= a \int_{-\infty}^{\infty} f_X(x) dx + b \int_{-\infty}^{\infty} x f_X(x) dx \\ \Rightarrow \quad \int_{-\infty}^{\infty} y \left[ \int_{-\infty}^{\infty} f(x,y) dx \right] dy &= a + bE(X) \\ \Rightarrow \quad \int_{-\infty}^{\infty} y f_Y(y) dy &= a + bE(X) \end{aligned}$$

$$\text{i.e., } E(Y) = a + bE(X) \Rightarrow \bar{Y} = a + b\bar{X} \quad \dots(3)$$

Multiplying both sides of (2) by  $x f_X(x)$  and integrating w.r.t.  $x$ , we get

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x,y) dy dx &= a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ \Rightarrow \quad E(XY) &= a E(X) + b E(X^2) \\ \Rightarrow \quad \mu_{11} + \bar{X} \bar{Y} &= a\bar{X} + b(\sigma_x^2 + \bar{X}^2) \end{aligned} \quad \dots(4)$$

$$\therefore \mu_{11} = E(XY) - E(X)E(Y) = E(XY) - \bar{X}\bar{Y};$$

$$\sigma_X^2 = E(X^2) - \{E(X)\}^2 = E(X^2) - \bar{X}^2$$

Solving (3) and (4) simultaneously, we get

$$b = \frac{\mu_{11}}{\sigma_X^2} \text{ and } a = \bar{Y} - \frac{\mu_{11}}{\sigma_X^2} \bar{X}$$

Substituting in (1) and simplifying, we get the required equation of the line of regression of  $Y$  on  $X$  as

$$\begin{aligned} E(Y|x) &= \bar{Y} + \frac{\mu_{11}}{\sigma_X^2} (x - \bar{X}) \\ \Rightarrow E(Y|X) &= \bar{Y} + \frac{\mu_{11}}{\sigma_X^2} (X - \bar{X}) \\ \Rightarrow E(Y|X) &= \bar{Y} + r \frac{\sigma_Y}{\sigma_X} (X - \bar{X}) \end{aligned}$$

By starting with the line  $E(X|y) = A + By$  and proceeding similarly we shall obtain the equation of the line of regression of  $X$  on  $Y$  as

$$E(X|Y) = \bar{X} + \frac{\mu_{11}}{\sigma_Y^2} (Y - \bar{Y}) = \bar{X} + r \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y})$$

### Example 10.19. Given

$$f(x, y) = xe^{-x(y+1)}, x \geq 0, y \geq 0,$$

find the regression curve of  $Y$  on  $X$ . [B.H. Univ. M.Sc., 1989]

**Solution.** Marginal p.d.f. of  $X$  is given by

$$\begin{aligned} f_1(x) &= \int_0^\infty f(x, y) dy = \int_0^\infty xe^{-x(y+1)} dy \\ &= xe^{-x} \int_0^\infty e^{-xy} dy = xe^{-x} \left[ \frac{e^{-xy}}{-x} \right]_0^\infty \\ &= e^{-x}, x \geq 0 \end{aligned}$$

Conditional p.d.f. of  $Y$  on  $X$  is given by

$$f(y|x) = \frac{f(x, y)}{f_1(x)} = \frac{xe^{-x(y+1)}}{e^{-x}} = xe^{-xy}, y \geq 0.$$

The regression curve of  $Y$  on  $X$  is given by

$$y = E(Y|X=x) = \int_0^\infty y f(y|x) dy = \int_0^\infty yxe^{-xy} dy$$

$$= x \left[ \left| \frac{ye^{-xy}}{-x} \right|_0^\infty + \int_0^\infty \frac{e^{-xy}}{x} dy \right] = 0 + \left| \frac{e^{-xy}}{-x} \right|_0^\infty = \frac{1}{x}$$

i.e.,  $y = \frac{1}{x} \Rightarrow xy = 1.$

which is the equation of a rectangular hyperbola. Hence the regression of  $Y$  on  $X$  is not linear.

**Example 10-20.** Obtain the regression equation of  $Y$  on  $X$  for the following distribution :

$$f(x, y) = \frac{y}{(1+x)^4} \exp\left(-\frac{y}{1+x}\right); x, y \geq 0$$

**Solution.** Marginal p.d.f. of  $X$  is given by

$$\begin{aligned} f_1(x) &= \int_0^\infty f(x, y) dy = \frac{1}{(1+x)^4} \int_0^\infty y e^{-y/(1+x)} dy \\ &= \frac{1}{(1+x)^4} \cdot \Gamma 2 \cdot (1+x)^2 \quad (\text{Using Gamma Integral}) \\ &= \frac{1}{(1+x)^2}; x \geq 0 \end{aligned}$$

The conditional p.d.f. of  $Y$  (for given  $X$ ) is

$$f(y|x) = \frac{f(x, y)}{f_1(x)} = \frac{y}{(1+x)^2} \exp\left(-\frac{y}{1+x}\right); y \geq 0$$

Regression equation of  $Y$  on  $X$  is given by

$$\begin{aligned} y &= E(Y|X) = \int_0^\infty y f(y|x) dy = \frac{1}{(1+x)^2} \int_0^\infty y^2 e^{-y/(1+x)} dy \\ &= \frac{1}{(1+x)^2} \cdot \Gamma 3 \cdot (1+x)^3 \quad [\text{Using Gamma Integral}] \\ \Rightarrow y &= 2(1+x) \quad [\because \Gamma 3 = 2! = 2] \end{aligned}$$

Hence the regression of  $Y$  on  $X$  is linear.

**Example 10-21.** Let  $(X, Y)$  have the joint p.d.f. given by

$$f(x, y) = \begin{cases} 1, & \text{if } |y| < x, 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Show that the regression of  $Y$  on  $X$  is linear but regression of  $X$  on  $Y$  is not linear.

**Solution.**  $|y| < x \Rightarrow -x < y < x$  and  $x > |y|$ .

The marginal p.d.f.'s  $f_1(\cdot)$  of  $X$  and  $f_2(\cdot)$  of  $Y$  are given by :

$$f_1(x) = \int_{-x}^x f(x, y) dy = \int_{-x}^x 1 dy = 2x; 0 < x < 1$$

$$f_2(y) = \int_{-|y|}^1 f(x, y) dx = \int_{-|y|}^1 1 dx = 1 - |y|; -1 < y < 1$$

$$\therefore f_1(x|y) = \frac{f(x,y)}{f_2(y)} = \frac{1}{1-|y|}; -1 \leq y < 1, 0 < x < 1$$

$$= \begin{cases} \frac{1}{1-y}, & 0 < y < 1; 0 < x < 1 \\ \frac{1}{1+y}, & -1 < y < 0; 0 < x < 1 \end{cases}$$

$$f_2(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{1}{2x}, 0 < x < 1; |y| < x$$

$$E(Y|X=x) = \int_{-x}^x y \cdot f_2(y|x) dy = \int_{-x}^x \frac{y}{2x} dy = \frac{1}{4x} \cdot |y^2| \Big|_{-x}^x = 0$$

Hence the curve of regression of  $Y$  on  $X$  is  $y = 0$ , which is a straight line.

$$E(X|Y=y) = \int x f_1(x|y) dx$$

$$\therefore E(X|Y=y) = \int_0^1 x \left( \frac{1}{1-y} \right) dx = \frac{1}{2(1-y)}, 0 < y < 1$$

$$\text{and } E(X|Y=y) = \int_0^1 x \left( \frac{1}{1+y} \right) dx = \frac{1}{2(1+y)}, -1 < y < 0$$

Hence the curve of regression of  $X$  on  $Y$  is

$$x = \begin{cases} \frac{1}{2(1-y)}, & 0 < y < 1 \\ \frac{1}{2(1+y)}, & -1 < y < 0, \end{cases}$$

which is not a straight line.

**Example 10-22.** Variables  $X$  and  $Y$  have the joint p.d.f.

$$f(x,y) = \frac{1}{3}(x+y), 0 \leq x \leq 1, 0 \leq y \leq 2.$$

*Find :*

- (i)  $r(X, Y)$
- (ii) The two lines of regression
- (iii) The two regression curves for the means.

**Solution.** The marginal p.d.f.'s of  $X$  and  $Y$  are given by :

$$f_1(x) = \int_0^2 f(x,y) dy = \frac{1}{3} \int_0^2 (x+y) dy = \frac{1}{3} \left| xy + \frac{y^2}{2} \right|_0^2$$

$$\Rightarrow f_1(x) = \frac{2}{3}(1+x); 0 \leq x \leq 1 \quad \dots(1)$$

$$f_2(y) = \int_0^1 f(x,y) dx = \frac{1}{3} \int_0^1 (x+y) dx = \frac{1}{3} \left| \frac{x^2}{2} + xy \right|_0^1$$

$$\Rightarrow f_2(y) = \frac{1}{3} \left( \frac{1}{2} + y \right); 0 \leq y \leq 2 \quad \dots(2)$$

The conditional distributions are given by :

$$f_3(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{1}{2} \left( \frac{x+y}{1+x} \right)$$

$$f_4(x|y) = \frac{f(x,y)}{f_2(y)} = \frac{2(x+y)}{1+2y} \quad . \quad \dots(3)$$

$$\begin{aligned} E(Y|x) &= \int_0^2 y \cdot f_3(y|x) dy = \frac{1}{2(1+x)} \int_0^2 y(x+y) dy \\ &= \frac{1}{2(1+x)} \left[ \frac{xy^2}{2} + \frac{y^3}{3} \right]_{y=0}^{y=2} = \frac{3x+4}{3(x+1)} \end{aligned}$$

Similarly, we shall get

$$E(X|y) = \int_0^1 x f_4(x|y) dx = \frac{2}{1+2y} \int_0^1 (x^2 + xy) dx = \frac{2+3y}{3(1+2y)}$$

(iii) Hence the regression curves for means are :

$$y = E(Y|x) = \frac{3x+4}{3(x+1)} \quad \text{and} \quad x = E(X|y) = \frac{2+3y}{3(1+2y)}.$$

From the marginal distributions we shall get

$$E(X) = \int_0^1 x f_1(x) dx = \frac{5}{9}, \quad E(X^2) = \int_0^1 x^2 f_1(x) dx = \frac{7}{18}$$

$$\Rightarrow \text{Var}(X) = \sigma_X^2 = \frac{7}{18} - \left(\frac{5}{9}\right)^2 = \frac{13}{162}$$

Similarly, we shall get

$$E(Y) = \frac{11}{9}, \quad E(Y^2) = \frac{16}{9}; \quad \sigma_Y^2 = \frac{16}{9} - \left(\frac{11}{9}\right)^2 = \frac{23}{81}$$

$$\begin{aligned} \text{Also } E(XY) &= \int_0^1 \int_0^2 xy f(x,y) dx dy = \frac{1}{3} \int_0^1 \int_0^2 (x^2 y + xy^2) dx dy \\ &= \frac{1}{3} \left\{ \left( \int_0^1 x^2 dx \right) \left( \int_0^2 y dy \right) + \left( \int_0^1 x dx \right) \left( \int_0^2 y^2 dy \right) \right\} \\ &= \frac{1}{3} \left[ \frac{1}{3} \times 2 + \frac{1}{2} \times \frac{8}{3} \right] = \frac{2}{3} \end{aligned}$$

$$\therefore \text{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{2}{3} - \frac{5}{9} \times \frac{11}{9} = -\frac{1}{81}$$

$$(i) \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{-\frac{1}{81}}{\sqrt{\frac{13}{162} \times \frac{23}{81}}} = -\left(\frac{2}{299}\right)^{1/2}$$

(ii) The two lines of regression are :

$$Y - E(Y) = \frac{\text{Cov}(X, Y)}{\sigma_Y^2} [X - E(X)] \Rightarrow Y - \frac{11}{9} = -\frac{2}{13} \left( X - \frac{5}{9} \right)$$

$$\text{and } X - E(X) = \frac{\text{Cov}(X, Y)}{\sigma_Y^2} [Y - E(Y)] \Rightarrow X - \frac{5}{9} = -\frac{1}{23} \left( Y - \frac{11}{9} \right)$$

**10.7.3. Regression Coefficients.** 'b', the slope of the line of regression of  $Y$  on  $X$  is also called the coefficient of regression of  $Y$  on  $X$ . It represents the increment in the value of dependent variable  $Y$  corresponding to a unit change in the value of independent variable  $X$ . More precisely, we write

$$b_{YX} = \text{Regression coefficient of } Y \text{ on } X = \frac{\mu_{11}}{\sigma_X^2} = r \frac{\sigma_Y}{\sigma_X} \quad \dots(10.17)$$

Similarly, the coefficient of regression of  $X$  on  $Y$  indicates the change in the value of variable  $X$  corresponding to a unit change in the value of variable  $Y$  and is given by

$$b_{XY} = \text{Regression coefficient of } X \text{ on } Y = \frac{\mu_{11}}{\sigma_Y^2} = r \frac{\sigma_X}{\sigma_Y} \quad \dots(10.17a)$$

#### 10.7.4. Properties of Regression Coefficients.

(a) Correlation coefficient is the geometric mean between the regression coefficients.

**Proof.** Multiplying (10.17) and (10.17a), we get

$$b_{XY} \times b_{YX} = r \frac{\sigma_X}{\sigma_Y} \times r \frac{\sigma_Y}{\sigma_X} = r^2$$

$$\therefore r = \pm \sqrt{b_{XY} \times b_{YX}} \quad \dots(10.18)$$

**Remark.** We have

$$r = \frac{\mu_{11}}{\sigma_X \cdot \sigma_Y}, \quad b_{YX} = \frac{\mu_{11}}{\sigma_X^2} \quad \text{and} \quad b_{XY} = \frac{\mu_{11}}{\sigma_Y^2}$$

It may be noted that the sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term  $\mu_{11}$ . Thus if the regression coefficients are positive, 'r' is positive and if the regression coefficients are negative 'r' is negative.

From (10.18), we have

$$r = \pm \sqrt{b_{XY} \times b_{YX}}$$

the sign to be taken before the square root is that of the regression coefficients.

(b) If one of the regression coefficients is greater than unity, the other must be less than unity.

**Proof.** Let one of the regression coefficients (say)  $b_{YX}$  be greater than unity, then we have to show that  $b_{XY} < 1$ .

$$\text{Now } b_{YX} > 1 \Rightarrow \frac{1}{b_{YX}} < 1 \quad \dots(*)$$

$$\text{Also } r^2 \leq 1 \Rightarrow b_{YX} \cdot b_{XY} \leq 1$$

$$\text{Hence } b_{XY} \leq \frac{1}{b_{YX}} < 1 \quad [\text{From } (*)]$$

(c) Arithmetic mean of the regression coefficients is greater than the correlation coefficient  $r$ , provided  $r > 0$ .

**Proof.** We have to prove that  $\frac{1}{2}(b_{YX} + b_{XY}) \geq r$

$$\text{or } \frac{1}{2} \left( r \frac{\sigma_Y}{\sigma_X} + r \frac{\sigma_X}{\sigma_Y} \right) \geq r \quad \text{or } \frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y} \geq 2 \quad (\because r > 0)$$

$$\Rightarrow \sigma_Y^2 + \sigma_X^2 - 2\sigma_X\sigma_Y \geq 0 \quad \text{i.e.,} \quad (\sigma_Y - \sigma_X)^2 \geq 0$$

which is always true, since the square of a real quantity is  $\geq 0$ .

(d) *Regression coefficients are independent of the change of origin but not of scale.*

**Proof.** Let  $U = \frac{X - a}{h}$ ,  $V = \frac{Y - b}{k} \Rightarrow X = a + hU$ ,  $Y = b + kV$ ,

where  $a, b, h (> 0)$  and  $k (> 0)$  are constants.

Then  $\text{Cov}(X, Y) = hk \text{Cov}(U, V)$ ,  $\sigma_X^2 = h^2\sigma_U^2$  and  $\sigma_Y^2 = k^2\sigma_V^2$

$$\begin{aligned} b_{YX} &= \frac{\mu_{11}}{\sigma_X^2} = \frac{hk \text{cov}(U, V)}{h^2\sigma_U^2} \\ &= \frac{k}{h} \cdot \frac{\text{cov}(U, V)}{\sigma_U^2} = \frac{k}{h} b_{UV} \end{aligned}$$

Similarly, we can prove that

$$b_{XY} = (h/k) b_{UV}$$

**10.7.5. Angle Between Two Lines of Regression.** Equations of the lines of regression of  $Y$  on  $X$ , and  $X$  on  $Y$  are

$$Y - \bar{y} = r \cdot \frac{\sigma_Y}{\sigma_X} (X - \bar{x}) \quad \text{and} \quad X - \bar{x} = r \cdot \frac{\sigma_X}{\sigma_Y} (Y - \bar{y})$$

Slopes of these lines are  $r \cdot \frac{\sigma_Y}{\sigma_X}$  and  $\frac{\sigma_Y}{r\sigma_X}$  respectively. If  $\theta$  is the angle between the two lines of regression then

$$\begin{aligned} \tan \theta &= \frac{r \frac{\sigma_Y}{\sigma_X} - \frac{\sigma_Y}{r\sigma_X}}{1 + r \frac{\sigma_Y}{\sigma_X} \cdot \frac{\sigma_Y}{r\sigma_X}} = \frac{r^2 - 1}{r} \left( \frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2} \right) \\ &= \frac{1 - r^2}{r} \left( \frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2} \right) \quad (\because r^2 \leq 1) \\ \therefore \theta &= \tan^{-1} \left\{ \frac{1 - r^2}{r} \left( \frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2} \right) \right\} \quad \dots (10.19) \end{aligned}$$

**Case (i).** ( $r = 0$ ). If  $r = 0$ ,  $\tan \theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

Thus if the two variables are uncorrelated, the lines of regression become perpendicular to each other.

**Case (ii).** ( $r = \pm 1$ ). If  $r = \pm 1$ ,  $\tan \theta = 0 \Rightarrow \theta = 0$  or  $\pi$ .

In this case the two lines of regression either coincide or they are parallel to each other. But since both the lines of regression pass through the point

( $\bar{x}$ ,  $\bar{y}$ ), they cannot be parallel. Hence in the case of perfect correlation, positive or negative, the two lines of regression coincide.

**Remarks 1.** Whenever two lines intersect, there are two angles between them, one acute angle and the other obtuse angle. Further  $\tan \theta > 0$  if  $0 < \theta < \pi/2$ , i.e.,  $\theta$  is an acute angle and  $\tan \theta < 0$  if  $\pi/2 < \theta < \pi$ , i.e.,  $\theta$  is an obtuse angle and since  $0 < r^2 < 1$ , the acute angle ( $\theta_1$ ) and obtuse angle  $\theta_2$  between the two lines of regression are given by

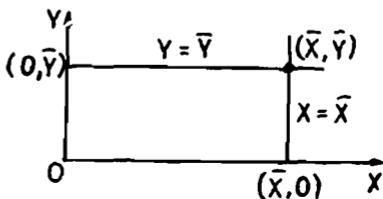
$$\theta_1 = \text{Acute angle} = \tan^{-1} \left\{ \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \cdot \frac{1 - r^2}{r} \right\}, r > 0$$

and  $\theta_2 = \text{Obtuse angle} = \tan^{-1} \left\{ \frac{\sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2} \cdot \frac{r^2 - 1}{r} \right\}, r > 0$

**2.** When  $r = 0$ , i.e., variables  $X$  and  $Y$  are uncorrelated, then the lines of regressions of  $Y$  on  $X$  and  $X$  on  $Y$  are given respectively by : [From (10-14a) and (10-15a)]

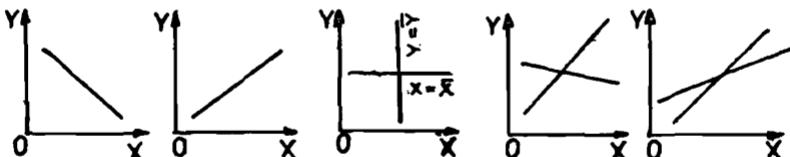
$$Y = \bar{Y} \text{ and } X = \bar{X},$$

as shown in the adjoining diagram. Hence, in this case ( $r = 0$ ), the lines of regression are perpendicular to each other and are parallel to  $X$ -axis and  $Y$ -axis respectively.



**3.** The fact that if  $r = 0$  (variables uncorrelated), the two lines of regression are perpendicular to each and if  $r = \pm 1$ ,  $\theta = 0$ , i.e., the two lines coincide, leads us to the conclusion that for higher degree of correlation between the variables, the angle between the lines is smaller, i.e., the two lines of regression are nearer to each other. On the other hand, if the lines of regression make a larger angle, they indicate a poor degree of correlation between the variables and ultimately for  $\theta = \pi/2$ ,  $r = 0$ , i.e., the lines become perpendicular if no correlation exists between the variables. Thus by plotting the lines of regression on a graph paper, we can have an approximate idea about the degree of correlation between the two variables under study. Consider the following illustrations :

TWO LINES COINCIDE ( $r = -1$ )	TWO LINES COINCIDE ( $r = +1$ )	TWO LINES PERPENDICULAR ( $r = 0$ )	TWO LINES APART (LOW DEGREE OF CORRELATION)	TWO LINES APART (HIGH DEGREE OF CORRELATION)
------------------------------------	------------------------------------	--	---	--



**10-7-6. Standard Error of Estimate or Residual Variance.** The equation of the line of regression of  $Y$  on  $X$  is

$$Y = \bar{Y} + r \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$$

$$\Rightarrow \frac{Y - \bar{Y}}{\sigma_Y} = r \left( \frac{X - \bar{X}}{\sigma_X} \right)$$

The residual variance  $s_Y^2$  is the expected value of the squares of deviations of the observed values of  $Y$  from the expected values as given by the line of regression of  $Y$  on  $X$ . Thus

$$\begin{aligned}s_Y^2 &= E[(Y - (\bar{Y} + r\sigma_Y(X - \bar{X})/\sigma_X))^2] \\&= \sigma_Y^2 E \left[ \frac{Y - \bar{Y}}{\sigma_Y} - r \left( \frac{X - \bar{X}}{\sigma_X} \right) \right]^2 = \sigma_Y^2 E(Y^* - rX^*)^2\end{aligned}$$

where  $X^*$  and  $Y^*$  are standardised variates so that

$$E(X^*^2) = 1 = E(Y^*^2) \text{ and } E(X^* Y^*) = r.$$

$$\therefore s_Y^2 = \sigma_Y^2 [E(Y^*^2) + r^2 E(X^*^2) - 2r E(X^* Y^*)] = \sigma_Y^2 (1 - r^2)$$

$$\Rightarrow s_Y = \sigma_Y (1 - r^2)^{1/2}$$

Similarly, the standard error of estimate of  $X$  is given by

$$s_X = \sigma_X (1 - r^2)^{1/2}$$

**Remarks 1.** Since  $s_X^2$  or  $s_Y^2 \geq 0$ , it follows that

$$(1 - r^2) \geq 0 \Rightarrow |r| \leq 1$$

Hence

$$-1 \leq r(X, Y) \leq 1$$

2. If  $r = \pm 1$ ,  $s_X = s_Y = 0$  so that each deviation is zero, and the two lines of regression are coincident.

3. Since, as  $r^2 \rightarrow 1$ ,  $s_X$  and  $s_Y \rightarrow 0$ , it follows that departure of the value  $r^2$  from unity indicates the departure of the relationship between the variables  $X$  and  $Y$  from linearity.

4. From the definition of linear regression, the minima condition implies that  $s_Y$  or  $s_X$  is the minimum variance.

**10.7.7. Correlation Coefficient between Observed and Estimated Value.** Here we will find the correlation between  $Y$  and

$$\hat{Y} = \bar{Y} + r \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$$

where  $\hat{Y}$  is the estimated value of  $Y$  as given by the line of regression of  $Y$  on  $X$ , which is given by

$$r(Y, \hat{Y}) = \frac{\text{Cov}(Y, \hat{Y})}{\sigma_Y \sigma_{\hat{Y}}}$$

We have

$$E(\hat{Y}) = E \left[ \bar{Y} + r \frac{\sigma_Y}{\sigma_X} (X - \bar{X}) \right] = \bar{Y} + r \frac{\sigma_Y}{\sigma_X} E(X - \bar{X}) = \bar{Y}$$

$$\therefore \text{Var}(\hat{Y}) = E[\hat{Y} - E(\hat{Y})]^2 = E \left[ r \frac{\sigma_Y}{\sigma_X} (X - \bar{X}) \right]^2 = r^2 \sigma_Y^2$$

$$\Rightarrow \hat{\sigma}_Y = r \sigma_Y$$

$$\text{Also } \text{Cov}(Y, \hat{Y}) = E[(Y - E(Y))( \hat{Y} - E(\hat{Y}))]$$

$$= E\left[\{b(X - E(X))\} \left\{r \frac{\sigma_Y}{\sigma_X} (X - \bar{X})\right\}\right]$$

$$= br \frac{\sigma_Y}{\sigma_X} E[(X - E(X))^2] = \left(r \frac{\sigma_Y}{\sigma_X}\right)^2 \sigma_X^2 = r^2 \sigma_Y^2$$

$$\therefore r(Y, \hat{Y}) = \frac{r^2 \sigma_Y^2}{\sigma_Y r \sigma_Y} = r = r(X, Y)$$

Hence the correlation coefficient between observed and estimated value of  $Y$  is the same as the correlation coefficient between  $X$  and  $Y$ .

**Example 10-23.** Obtain the equations of the lines of regression for the data in Example 10-1. Also obtain the estimate of  $X$  for  $\hat{Y} = 70$ .

**Solution.** Let  $U = X - 68$  and  $V = Y - 69$ , then

$$\bar{U} = 0, \bar{V} = 0, \sigma_U^2 = 4.5, \sigma_V^2 = 5.5, \text{Cov}(U, V) = 3 \text{ and } r(U, V) = 0.6$$

Since correlation coefficient is independent of change of origin, we get

$$r = r(X, Y) = r(U, V) = 0.6$$

We know that if  $U = \frac{X - a}{h}$ ,  $V = \frac{Y - b}{k}$ , then

$$\bar{X} = a + h\bar{U}, \bar{Y} = b + k\bar{V}, \sigma_X = h\sigma_U \text{ and } \sigma_Y = k\sigma_V$$

In our case  $h = k = 1$ ,  $a = 68$  and  $b = 69$ .

$$\text{Thus } \bar{X} = 68 + 0 = 68, \bar{Y} = 69 + 0 = 69$$

$$\sigma_X = \sigma_U = \sqrt{4.5} = 2.12 \text{ and } \sigma_Y = \sigma_V = \sqrt{5.5} = 2.35$$

Equation of line of regression of  $\hat{Y}$  on  $X$  is

$$Y - \bar{Y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$$

$$\text{i.e., } Y = 69 + 0.6 \times \frac{2.35}{2.12} (X - 68) \Rightarrow Y = 0.665 X + 23.78$$

Equation of line of regression of  $X$  on  $Y$  is

$$X - \bar{X} = r \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y})$$

$$\Rightarrow X = 68 + 0.6 \times \frac{2.12}{2.35} (Y - 69) \text{ i.e., } X = 0.54Y + 30.74$$

To estimate  $X$  for given  $Y$ , we use the line of regression of  $X$  on  $Y$ . If  $Y = 70$ , estimated value of  $X$  is given by

$$\hat{X} = 0.54 \times 70 + 30.74 = 68.54,$$

where  $\hat{X}$  is estimate of  $X$ .

**Example 10-24.** In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible :

Variance of  $X = 9$ .

Regression equations :  $8X - 10Y + 66 = 0$ ,  $40X - 18Y = 214$ .

What were (i) the mean values of  $X$  and  $Y$ ,

(ii) the correlation coefficient between  $X$  and  $Y$ , and

(iii) the standard deviation of  $Y$ ?

[Punjab Univ. B.Sc. (Hons.), 1993]

**Solution** (i) Since both the lines of regression pass through the point  $(\bar{X}, \bar{Y})$ , we have  $8\bar{X} - 10\bar{Y} + 66 = 0$ , and  $40\bar{X} - 18\bar{Y} = 214$ .

Solving, we get  $\bar{X} = 13$ ,  $\bar{Y} = 17$ .

(ii) Let  $8X - 10Y + 66 = 0$  and  $40X - 18Y = 214$  be the lines of regression of  $Y$  on  $X$  and  $X$  on  $Y$  respectively. These equations can be put in the form :

$$Y = \frac{8}{10}X + \frac{66}{10} \text{ and } X = \frac{18}{40}Y + \frac{214}{40}$$

$$\therefore b_{YX} = \text{Regression coefficient of } Y \text{ on } X = \frac{8}{10} = \frac{4}{5}$$

$$\text{and } b_{XY} = \text{Regression coefficient of } X \text{ on } Y = \frac{18}{40} = \frac{9}{20}$$

$$\text{Hence } r^2 = b_{YX} \cdot b_{XY} = \frac{4}{5} \cdot \frac{9}{20} = \frac{9}{25}$$

$$\therefore r = \pm \frac{3}{5} = \pm 0.6$$

But since both the regression coefficients are positive, we take  $r = +0.6$

$$(iii) \text{ We have } b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X} \Rightarrow \frac{4}{5} = \frac{3}{5} \times \frac{\sigma_Y}{3} [\because \sigma_X^2 = 9 \text{ (Given)}]$$

$$\text{Hence } \sigma_Y = 4$$

**Remarks.** 1. It can be verified that the values of  $\bar{X} = 13$  and  $\bar{Y} = 17$  as obtained in part (i) satisfy both the regression equations. In numerical problems of this type, this check should invariably be applied to ascertain the correctness of the answer.

2. If we had assumed that  $8X - 10Y + 66 = 0$ , is the equation of the line of regression of  $X$  on  $Y$  and  $40X - 18Y = 214$  is the equation of line of regression of  $Y$  on  $X$ , then we get respectively :

$$8X = 10Y - 66 \text{ and } 18Y = 40X - 214$$

$$\Rightarrow X = \frac{10}{8}Y - \frac{66}{8} \text{ and } Y = \frac{40}{18}X - \frac{214}{18}$$

$$\Rightarrow b_{XY} = \frac{18}{8} \text{ and } b_{YX} = \frac{40}{18}$$

$$\therefore r^2 = b_{XY} \cdot b_{YX} = \frac{10}{8} \times \frac{40}{18} = 2.78$$

But since  $r^2$  always lies between 0 and 1, our supposition is wrong.

**Example 10-25.** Find the most likely price in Bombay corresponding to the price of Rs. 70 at Calcutta from the following :

	Calcutta	Bombay
Average price	65	67
Standard deviation	2.5	3.5

Correlation coefficient between the prices of commodities in the two cities is 0.8 : [Nagpur Univ. B.Sc., 1993; Sri Venkateswara Univ. B.Sc. (Oct.) 1990]

**Solution.** Let the prices, (in Rupees), in Bombay and Calcutta be denoted by  $Y$  and  $X$  respectively. Then we are given

$\bar{X} = 65$ ,  $\bar{Y} = 67$ ,  $\sigma_X = 2.5$ ,  $\sigma_Y = 3.5$  and  $r = r(X, Y) = 0.8$ . We want  $Y$  for  $X = 70$ .

Line of regression of  $Y$  on  $X$  is

$$Y - \bar{Y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$$

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

$$\text{When } X = 70, \quad \hat{Y} = 67 + 0.8 \times \frac{3.5}{2.5} (70 - 65) = 72.6$$

**Example 10-26.** Can  $Y = 5 + 2.8X$  and  $X = 3 - 0.5Y$  be the estimated regression equations of  $Y$  on  $X$  and  $X$  on  $Y$  respectively ? Explain your answer with suitable theoretical arguments. [Delhi Univ. M.A.(Eco.), 1986]

**Solution.** Line of regression of  $Y$  on  $X$  is :

$$Y = 5 + 2.8X \Rightarrow b_{YX} = 2.8 \quad \dots(*)$$

Line of regression of  $X$  on  $Y$  is :

$$X = 3 - 0.5Y \Rightarrow b_{XY} = -0.5 \quad \dots(**)$$

This is not possible, since each of the regression coefficients  $b_{YX}$  and  $b_{XY}$  must have the same sign, which is same as that of  $\text{Cov}(X, Y)$ . If  $\text{Cov}(x, y)$  is positive, then both the regression coefficients are positive and if  $\text{Cov}(X, Y)$  is negative, then both the regression coefficients are negative. Hence (\*) and (\*\*) cannot be the estimated regression equations of  $Y$  on  $X$  and  $X$  on  $Y$  respectively.

### EXERCISE 10 (d)

1. (a) Explain what are regression lines. Why are there two such lines ? Also derive their equations.

(b) Define (i) Line of regression, (ii) Regression coefficient. Find the equations to the lines of regression and show that the coefficient of correlation is the geometric mean of coefficients of regression.

(c) What equation is the equivalent mathematical statement for the following words ?

"If the respective deviations in each series,  $X$  and  $Y$ , from their means were expressed in units of standard deviations, i.e., if each were divided by the

standard deviation of the series; to which it belongs and plotted to a scale of standard deviations, the slope of a straight line best describing the plotted points would be the correlation coefficient  $r$ .

2(a) Obtain the equation of the line of regression of  $Y$  on  $X$  and show that the angle  $\theta$ , between the two lines of regression is given by

$$\tan \theta = \frac{1 - \rho^2}{\rho} \times \frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}$$

where  $\sigma_1, \sigma_2$  are the standard deviations of  $X$  and  $Y$  respectively, and  $\rho$  is the correlation coefficient. (Delhi Univ. B.Sc. (Maths. Hons.), 1989)

Interpret the cases when  $\rho = 0$  and  $\rho = \pm 1$ .

(Bangalore Univ. B.Sc. 1990)

(b) If  $\theta$  is the acute angle between the two regression lines with correlation coefficient  $r$ , show that  $\sin \theta \leq 1 - r^2$ .

3. (a) Explain the term "regression" by giving examples. Assuming that the regression of  $Y$  on  $X$  is linear, outline a method for the estimation of the coefficients in the regression line based on the random paired sample of  $X$  and  $Y$ , and show that the variance of the error of the estimate for  $Y$  for the regression line is  $\sigma_Y^2(1 - \rho^2)$ , where  $\sigma_Y^2$  is the variance of  $Y$  and  $\rho$  is the correlation coefficient between  $X$  and  $Y$ .

(b) Prove that  $X$  and  $Y$  are linearly related if and only if  $\rho_{XY}^2 = 1$ . Further show that the slope of the regression line is positive or negative according as  $\rho = +1$  or  $\rho = -1$ .

(c) Let  $X$  and  $Y$  be two variates. Define  $X^* = \frac{X - a}{b}$ ,  $Y^* = \frac{Y - c}{d}$  for some constants  $a, b, c$  and  $d$ . Show that the regression line (least square) of  $Y$  on  $X$  can be obtained from that of  $Y^*$  on  $X^*$ .

(d) Show that the coefficient of correlation between the observed and the estimated values of  $Y$  obtained from the line of regression of  $Y$  on  $X$ , is the same as that between  $X$  and  $Y$ .

4. Two variables  $X$  and  $Y$  are known to be related to each other by the relation  $Y = X/(aX + b)$ . How is the theory of linear regression to be employed to estimate the constants  $a$  and  $b$  from a set of  $n$  pairs of observations  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ ?

Hint.

$$\frac{1}{Y} = \frac{aX + b}{X} = a + \frac{b}{X}$$

Put

$$\frac{1}{X} = U \text{ and } \frac{1}{Y} = V$$

$\therefore$

$$V = a + bU$$

5. Derive the standard error of estimate of  $Y$  obtained from the linear regression equation of  $Y$  on  $X$ . What does this standard error measure?

6. (a) Calculate the coefficient of correlation from the following data :

$X$ :	1	2	3	4	5	6	7	8	9
$Y$ :	9	8	10	12	11	13	14	16	15

Also obtain the equations of the lines of regression and obtain an estimate of  $Y$  which should correspond on the average to  $X = 6.2$ .

Ans.  $r = 0.95$ ,  $Y - 12 = 0.95(X - 5)$ ,  $X - 5 = 0.95(Y - 12)$ ,  $13.14$

(b) Why do we have, in general, two lines of regression ? Obtain the regression of  $Y$  on  $X$ , and  $X$  on  $Y$  from the following table and estimate the blood pressure when the age is 45 years :

<i>Age in years</i> (X)	<i>Blood pressure</i> (Y)	<i>Age in years</i> (X)	<i>Blood pressure</i> (Y)
56	147	55	150
42	125	49	145
72	160	38	115
36	118	42	140
63	149	68	152
47	128	60	155

Ans.  $Y = 1.138X + 80.778$ ,  $Y = 131.988$  for  $X = 45$ .

(c) Suppose the observations on  $X$  and  $Y$  are given as :

X : 59	65	45	52	60	62	70	55	45	49
Y : 75	70	55	65	60	69	80	65	59	61

where  $N = 10$  students, and  $Y$  = Marks in Maths,  $X$  = Marks in Economics. Compute the least square regression equations of  $Y$  on  $X$  and of  $X$  on  $Y$ .

If a student gets 61 marks in Economics, what would you estimate his marks in Maths to be ?

7. (a) In a correlation analysis on the ages of wives and husbands, the following data were obtained. Find

(i) the value of the correlation coefficient, and (ii) the lines of regression.

Estimate the age of husband whose wife's age is 31 years. Estimate the age of wife whose husband is 40 years old.

		Age of wife →	15—25	25—35	35—45	45—55	55—65
Age of Husband ↓	15—30	30	6	3	—	—	
	30—45	18	32	15	12	8	
	45—60	2	28	40	16	9	
	60—75	—	4	9	10	8	

(b) The following table gives the distribution of total cultivable area ( $X$ ) and area under cultivation ( $Y$ ) in a district of 69 villages.

Calculate (i) the linear regression of  $Y$  on  $X$ ,

(ii) the correlation coefficient  $r(X, Y)$ , and (iii) the average area under wheat corresponding to total area of 1,000 Bighas.

		Total area (in Bighas)				
		0—500	500—1000	1000—1500	1500—2000	2000—2500
Area under wheat	0—200	12	6	...	...	...
	200—400	2	18	4	2	1
	400—600	...	4	7	3	...
	600—800	...	1	...	2	1
	800—1000	...	...	1	2	3

$$\text{Ans. (i)} \quad Y = 0.7641X - 455.3854, \quad \text{(ii)} \quad r(X, Y) = 0.756$$

$$\text{(iii)} \quad Y = 308.7146 \text{ for } X = 1000$$

8. (a) Compare and contrast the roles of correlation and regression in studying the inter-dependence of two variates.

For 10 observations on price ( $X$ ) and supply ( $Y$ ) the following data were obtained (in appropriate units).

$$\Sigma X = 130, \Sigma Y = 220, \Sigma X^2 = 2288, \Sigma Y^2 = 5506 \text{ and } \Sigma XY = 3467$$

Obtain the line of regression of  $Y$  on  $X$  and estimate the supply when the price is 16 units, and find out the standard error of the estimate.

$$\text{Ans. } Y = 8.8 + 1.015X, 25.04$$

(b) If a number  $X$  is chosen at random from among the integers 1, 2, 3, 4 and a number  $Y$  is chosen from among those at least as large as  $X$ , prove that

$$\text{Cov}(X, Y) = \frac{5}{8}$$

Find also the regression line of  $X$  on  $Y$ .

(c) Calculate the correlation coefficient from the following data :—

$$N = 100, \quad \Sigma X = 12500 \quad \Sigma Y = 8000$$

$$\Sigma X^2 = 1585000, \quad \Sigma Y^2 = 648100 \quad \Sigma XY = 1007425.$$

Also obtain the regression equation of  $Y$  on  $X$ .

9. (a) The means of a bivariate frequency distribution are at (3, 4) and  $r = 0.4$ . The line of regression of  $Y$  on  $X$  is parallel to the line  $Y = X$ . Find the two lines of regression and estimate the mean of  $X$  when  $Y = 1$ .

(b) For certain data,  $Y = 1.2 X$  and  $X = 0.6 Y$ , are the regression lines. Compute  $\rho(X, Y)$  and  $\sigma_X/\sigma_Y$ . Also compute  $\rho(X, Z)$ , if  $Z = Y - X$ .

(c) The equations of two regression lines obtained in a correlation analysis are as follows :

$$3X + 12Y = 19, \quad 3Y + 9X = 46$$

Obtain (i) the value of correlation coefficient,

(ii) mean values of  $X$  and  $Y$ , and

(iii) the ratio of the coefficient of variability of  $X$  to that of  $Y$ .

Ans. (i)  $-\frac{1}{2}\sqrt{3}$ , (ii)  $\bar{X} = 5$ ,  $\bar{Y} = 1/3$ .

(d) For an army personnel of strength 25, the regression of weight of kidneys ( $Y$ ) on weight of heart ( $X$ ), both measured in ounces is

$$Y - 0.399X - 6.934 = 0$$

and the regression of weight of heart on weight of kidney is

$$X - 1.212Y + 2.461 = 0$$

Find the correlation coefficient between  $X$  and  $Y$  and their mean values. Can you find out the standard deviation of  $X$  and  $Y$  as well?

Ans.  $r(X, Y) = 0.70$ ,  $\bar{X} = 11.5086$ ,  $\bar{Y} = 11.5261$ , No.

(e) Find the coefficient of correlation for distribution in which

$$\text{S.D. of } X = 3.0 \text{ units}$$

$$\text{S.D. of } Y = 1.4 \text{ units}$$

Coefficient of regression of  $Y$  on  $X = 0.28$ .

10. (a) Given that  $X = 4Y + 5$  and  $Y = kX + 4$ , are the lines of regression of  $X$  on  $Y$  and  $Y$  on  $X$  respectively, show that  $0 < 4k < 1$ . If  $k = \frac{1}{16}$ , find the means of the two variables and coefficient of correlation between them.

[Punjab Univ. B.Sc. (Hons.), 1989]

Hint.  $X = 4Y + 5 \Rightarrow b_{XY} = 4$

$$Y = kx + 4 \Rightarrow b_{YX} = k$$

$$\therefore r^2 = 4k \quad \dots(*)$$

$$\text{But } 0 \leq r^2 \leq 1 \Rightarrow 0 \leq 4k \leq 1.$$

$$\text{If } k = \frac{1}{16}, \text{ then from } (*), \text{ we get}$$

$$r^2 = 4 \times \frac{1}{16} \Rightarrow r = +\frac{1}{2} \quad [\text{Since both the regression coefficient are positive}]$$

$$\text{For } k = \frac{1}{16}, \text{ the two lines of regression become}$$

$$X = 4Y + 5 \text{ and } Y = \frac{1}{16}X + 4$$

Solving the two equations, we get  $\bar{Y} = 5.75$ ,  $\bar{X} = 28$ .

(b) For 50 students of a class the regression equation of marks in Statistics ( $X$ ) on marks in Mathematics ( $Y$ ) is  $3Y - 5X + 180 = 0$ . The mean marks in Mathematics is 44 and variance of marks in Statistics is  $9/16$ th of the variance of marks in Mathematics. Find the mean marks in Statistics and the coefficient of correlation between marks in two subjects.

[Bangalore Univ. B.Sc., 1989]

Hint. We are given  $n = 50$ ,  $\bar{Y} = 44$

$$\text{and } \sigma_X^2 = \frac{9}{16} \sigma_Y^2 \Rightarrow \frac{\sigma_X}{\sigma_Y} = \frac{3}{4} \quad \dots(*)$$

The equation of the line of regression of  $X$  on  $Y$  is given to be

$$3Y - 5X + 180 = 0 \Rightarrow X = \frac{3}{5}Y + \frac{180}{5}$$

$$\therefore b_{XY} = r \frac{\sigma_X}{\sigma_Y} = \frac{3}{5} \Rightarrow r \cdot \frac{3}{4} = \frac{3}{5} \quad \text{or} \quad r = 0.8$$

Since the lines of regression pass through the point  $(\bar{X}, \bar{Y})$ , we get

$$\bar{X} = \frac{3}{5}\bar{Y} + \frac{180}{5} = \frac{3}{5} \times 44 + 36 = 62.4$$

(c) Out of the two lines of regression given by

$$X + 2Y - 5 = 0 \text{ and } 2X + 3Y - 8 = 0,$$

which one is the regression line of  $X$  on  $Y$ ?

Use the equations to find the mean of  $X$  and the mean of  $Y$ . If the variance of  $X$  is 12, calculate the variance of  $Y$ .

Ans.  $\bar{X} = 1, \bar{Y} = 2, \sigma_Y^2 = 4$

(d) The lines of regression in a bivariate distribution are :

$$X + 9Y = 7 \text{ and } Y + 4X = \frac{49}{3}$$

Find (i) the coefficient of correlation, (iii) the ratios  $\sigma_X^2 : \sigma_Y^2 : \text{Cov}(X, Y)$ , (iii) the means of the distribution and (iv)  $E(X | Y = 1)$ .

(e) Estimate  $X$  when  $Y = 10$ , if the two lines of regression are :

$$X = -\frac{1}{18}Y + \lambda \text{ and } Y = -2x + \mu,$$

$(\lambda, \mu)$  being unknown and the mean of the distribution is at  $(-1, 2)$ . Also compute  $r, \lambda$  and  $\mu$ . [Gujarat Univ. B.Sc., Oct. 1992]

11. (a) The following results were obtained in the analysis of data on yield of dry bark in ounces ( $Y$ ) and age in years ( $X$ ) of 200 cinchóna plants :

	X	Y
Average	9.2	16.5
Standard deviation	2.1	4.2
Correlation coefficient	+0.84	

Construct the two lines of regression and estimate the yield of dry bark of a plant of age 8 years. [Patna Univ. B.Sc., 1991],

(b) The following data pertain to the marks in subjects  $A$  and  $B$  in a certain examination :

Mean marks in  $A = 39.5$

Mean marks in  $B = 47.5$

Standard deviation of marks in  $A = 10.8$

Standard deviation of marks in  $B = 16.8$

Coefficient of correlation between marks in  $A$  and marks in  $B = 0.42$ .

Draw the two lines of regression and explain why there are two regression equations. Give the estimate of marks in  $B$  for candidates who secured 50 marks in  $A$ .

Ans.  $Y = 0.65X + 21.825, X = 0.27Y + 26.675$  and  $Y = 54.342$  for  $X = 50$

(c) You are given the following information about advertising expenditure and sales :

	<i>Advertising Expenditure (X)</i> (Rs. lakhs)	<i>Sales (Y)</i> (Rs. lakhs)
Mean	10	90
s.d.	3	12

Correlation coefficient = 0.8

What should be the advertising budget if the company wants to attain sales target of Rs. 120 lakhs ? [Delhi Univ. M.C.A., 1990]

12. Twenty-five pairs of value of variates  $X$  and  $Y$  led to the following results :

$$N = 25, \sum X = 127, \sum Y = 100, \sum X^2 = 760, \sum Y^2 = 449 \text{ and } \sum XY = 500$$

A subsequent scrutiny showed that two pairs of values were copied down as :

<i>X</i>	<i>Y</i>
8	14
8	6

<i>X</i>	<i>Y</i>
8	12
6	8

(i) Obtain the correct value of the correlation coefficient.

(ii) Hence or otherwise, find the correct equations of the two lines of regression.

(iii) Find the angle between the regression lines.

Ans. (i)  $r(X, Y) = -(0.64 \times 0.15)^{1/2}$ ,

$$(ii) X = -0.64Y + 7.56, Y = -0.15X + 4.75.$$

13. Suppose you have  $n$  observations :

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

on two variables  $X$  and  $Y$ , and you have fitted a linear regression  $Y = a + bX$  by the method of least squares. Denote the 'expected' value of  $Y$  by  $Y^*$ , and the residual  $Y - Y^*$  by  $e$ . Find means and variances of  $Y^*$  and  $e$ , and the correlation co-efficient between (i)  $X$  and  $e$ , (ii)  $Y$  and  $e$  and (iii)  $Y$  and  $Y^*$ . Use these results to bring out the significance and limitations of the correlation coefficient.

Ans.  $r(X, e) = 0, r(Y, e) = 0$  and  $r(Y, Y^*) = r(X, Y)$ .

14. (a) The regression lines of  $Y$  on  $X$  and of  $X$  on  $Y$  are respectively  $Y = aX + b$  and  $X = cY + d$ . Show that

(i) Means are  $\bar{X} = (bc + d)/(1 - ac)$  and  $\bar{Y} = (ad + b)/(1 - ac)$

(ii) Correlation coefficient between  $X$  and  $Y$  is  $\sqrt{ac}$ .

(iii) The ratio of the standard deviations of  $X$  and  $Y$  is  $\sqrt{c/a}$ .

(b) For two random variables  $X$  and  $Y$  with the same mean, the two regression equations are  $Y = aX + b$  and  $X = \alpha Y + \beta$ . Show that  $\frac{b}{\beta} = \frac{1 - \alpha}{1 - \alpha}$ . Find also the common mean.

[Punjab Univ. B.Sc. (Maths Hons.), 1992]

(c) If the lines of regression of  $Y$  on  $X$  and  $X$  on  $Y$  are respectively  $a_1X + b_1Y + c_1 = 0$  and  $a_2X + b_2Y + c_2 = 0$ , prove that  $a_1b_2 \leq a_2b_1$ .

(Delhi Univ. B.Sc. (Stat. Hons.), 1989)

$$\text{Hint. } r^2 = b_{YX} \cdot b_{XY} \leq 1 \Rightarrow \left( -\frac{a_1}{b_1} \right) \times \left( -\frac{b_2}{a_2} \right) = \frac{a_1 b_2}{a_2 b_1} \leq 1$$

15. (a) By minimising  $\sum_{i=1}^n f_i (x_i \cos \alpha + y_i \sin \alpha - p)^2$  for variations in  $\alpha$

and  $p$ , show that there are two straight lines passing through the mean of the distribution for which the sum of squares of normal deviations has an extreme value. Prove also that their slopes are given by

$$\tan 2\alpha = \frac{2\mu_{11}}{\sigma_x^2 - \sigma_y^2}$$

**Hint.** We have to minimize

$$S = \sum_{i=1}^n f_i (x_i \cos \alpha + y_i \sin \alpha - p)^2 \quad \dots(1)$$

Equating to zero, the partial derivatives of (1) w.r.t.  $\alpha$  and  $p$ , we have

$$\frac{\partial S}{\partial \alpha} = 0 = 2 \sum_{i=1}^n f_i (x_i \cos \alpha + y_i \sin \alpha - p) (-x_i \sin \alpha + y_i \cos \alpha) \quad \dots(2)$$

$$\frac{\partial S}{\partial p} = 0 = -2 \sum_{i=1}^n f_i (x_i \cos \alpha + y_i \sin \alpha - p) \quad \dots(3)$$

Equation (3) can be written as

$$\sum_{i=1}^n f_i (x_i \cos \alpha + y_i \sin \alpha - p) = 0 \Rightarrow \bar{x} \cos \alpha + \bar{y} \sin \alpha - p = 0 \quad \dots(4)$$

From equation (2), we get a quadratic equation which shows that there are two straight lines for extreme values of  $E$ .

From equation (4), it becomes clear that both the straight lines pass through the point  $(\bar{x}, \bar{y})$ .

Again equation (2) can be written as :

$$\begin{aligned} & \sum_{i=1}^n f_i (x_i \cos \alpha + y_i \sin \alpha - p) (y_i \cos \alpha - x_i \sin \alpha) = 0 \\ \Rightarrow & \sum_{i=1}^n f_i [\cos \alpha (x_i - \bar{x}) + \sin \alpha (y_i - \bar{y})] [y_i \cos \alpha - x_i \sin \alpha] = 0 \\ & \quad [\text{Using (4)}] \\ \Rightarrow & \cos^2 \alpha \sum_{i=1}^n f_i y_i (x_i - \bar{x}) - \sin \alpha \cos \alpha \sum_{i=1}^n f_i x_i (x_i - \bar{x}) \\ & + \sin \alpha \cos \alpha \sum_{i=1}^n f_i y_i (y_i - \bar{y}) - \sin^2 \alpha \sum_{i=1}^n f_i x_i (y_i - \bar{y}) = 0 \quad \dots(5) \end{aligned}$$

$$\text{We have } \mu_{11} = \frac{1}{N} \sum_i f_i (x_i - \bar{x}) (y_i - \bar{y})$$

$$= \frac{1}{N} \sum_i f_i x_i (y_i - \bar{y}) - \bar{x} \cdot \frac{1}{N} \sum_i f_i (y_i - \bar{y}) = \frac{1}{N} \sum_i f_i x_i (y_i - \bar{y})$$

Similarly,  $\mu_{11} = \frac{1}{N} \sum_i f_i y_i (x_i - \bar{x})$

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i x_i (x_i - \bar{x}) \text{ and } \sigma_y^2 = \frac{1}{N} \sum_i f_i y_i (y_i - \bar{y})$$

Substituting these values in (5), we get the required result.

(b) If the straight line defined by

$$Y = a + bX$$

satisfies the condition  $E[(Y - a - bX)^2] = \text{minimum}$ , show that the regression line of the random variable  $Y$  on the random variable  $X$  is

$$Y - \bar{Y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{X}), \text{ where } \bar{X} = E(X), \bar{Y} = E(Y)$$

16. (a) Define Curve of regression of  $Y$  on  $X$ .

The joint density function of  $X$  and  $Y$  is given by :

$$\begin{aligned} f(x, y) &= x + y, 0 < x < 1, 0 < y < 1 \\ &= 0, \text{ otherwise} \end{aligned}$$

Find

- (i) the correlation coefficient between  $X$  and  $Y$ ,
- (ii) the regression curve of  $Y$  on  $X$ , and
- (iii) the regression curve of  $X$  on  $Y$ .

Ans.  $\rho(X, Y) = -\frac{1}{11}$ . [Madras Univ. B.Sc., Stat. (Main), 1992]

$$(b) \text{ Let } f(x_1, x_2) = \frac{2}{a^2}; 0 < x_1 < x_2, 0 < x_2 < a \\ = 0, \text{ elsewhere}$$

be the joint p.d.f. of  $X_1$  and  $X_2$ .

Find conditional means and variances. Also show that  $\rho = \frac{1}{2}$ .

17. If the joint density of  $X$  and  $Y$  is given by

$$f(x, y) = \begin{cases} (x + y)/3, & \text{for } 0 < x < 1, 0 < y < 2 \\ 0, & \text{otherwise} \end{cases}$$

obtain the regressions (i) of  $Y$  on  $X$  and (ii) of  $X$  on  $Y$ .

Are the regressions linear? Find the correlation coefficient between  $X$  and  $Y$ . (Allahabad Univ. B.Sc. 1992)

$$\text{Ans. } y = E(Y|x) = \frac{3x + 4}{3(x + 1)}; x = E(X|y) = \frac{2 + 3y}{3(1 + 2y)}$$

$$\text{Corr. } (X, Y) = -\left(\frac{2}{299}\right)^{1/2}$$

18. Let the joint density function of  $X$  and  $Y$  be given by

$$\begin{aligned} f(x, y) &= 8xy, 0 < x < y < 1 \\ &= 0, \text{ otherwise} \end{aligned}$$

Find: (i)  $E(Y|X=x)$ , (ii)  $E[XY|X=x]$ , (iii)  $\text{Var}[Y|X=x]$

[Delhi Univ. B.Sc. (Maths Hons.), 1988]

Ans. (i)  $E(Y|x) = \frac{2}{3} \left( \frac{1+x+x^2}{1+x} \right)$ ;  $E(XY|x) = x E(Y|x)$ , (iii)  $E(Y^2|x) = \frac{1+x^2}{2}$

19. Give an example to show that it is possible to have the regression of  $Y$  on  $X$  constant (does not depend on  $X$ ), but the regression of  $X$  on  $Y$  is not constant (does depend on  $Y$ ).

Hint. See Example 10.21.

20. Prove or disprove :

$$E(Y|X=x) = \text{constant} \Rightarrow r(X,Y) = 0$$

Ans. True

21. If  $f(x,y) = \frac{1}{3}x^2 \exp[-y(1+x)]$ ,  $x \geq 0, y \geq 0$ , is the joint p.d.f. of  $(X,Y)$ , obtain the equation of regression of  $Y$  on  $X$ .

Ans.  $y = E(Y|x) = 1/(1+x)$ .

22. Variables  $(X,Y)$  have joint p.d.f.

$$f(x,y) = 6(1-x-y), x > 0, y > 0, x+y < 1, \\ = 0, \text{ otherwise.}$$

Find  $f_X(x), f_Y(y)$  and  $\text{Cov}(X,Y)$ . Are  $X$  and  $Y$  independent? Obtain the regression curves for the means.

[Calcutta Univ. B.Sc. (Maths Hons.), 1986]

Ans.  $f_1(x) = 3(1-x)^2, 0 < x < 1; f_2(y) = 3(1-y)^2, 0 < y < 1$ .

$X$  and  $Y$  are not independent.

Regression curves for the means are:

$$y = E(Y|x) = \frac{1}{3}(1-x) \text{ and } x = E(X|y) = \frac{1}{3}(1-y).$$

23. For the joint p.d.f.

$$f(x,y) = 3x^2 - 8xy + 6y^2, 0 \leq x, y \leq 1,$$

find the least square regression lines and the regression curves for the means.

[Calcutta Univ. B.Sc. (Maths, Hons.), 1987]

Ans. Regression lines :

$$y - \frac{2}{3} = -\frac{10}{67}\left(x - \frac{5}{12}\right); \quad x - \frac{5}{12} = -\frac{25}{32}\left(y - \frac{2}{3}\right)$$

Regression curves for means are :

$$y = E(Y|x) = \frac{9x^2 - 16x + 9}{6(3x^2 - 4x + 2)}; \quad x = E(X|y) = \frac{36y^2 - 32y - 9}{12(6y^2 - 4y + 1)}$$

24. Let  $(X, Y)$  be jointly distributed with p.d.f.

$$f(x,y) = e^{-y}, 0 < x < y < \infty$$

$$= 0, \text{ otherwise}$$

Prove that :

$$E(Y|X=x) = x + \bar{Y} \text{ and } E(X|Y=y) = y/2.$$

Hence prove that  $r(X, Y) = \sqrt{1/2}$ .

25. Let  $f(x, y) = e^{-y} (1 - e^{-x})$ ,  $0 < x < y ; 0 < y < \infty$   
 $= e^{-x} (1 - e^{-y})$ ,  $0 < y < x ; 0 < x < \infty$

- (a) Show that  $f(x, y)$  is a p.d.f.
- (b) Find marginal distributions of  $X$  and  $Y$ .
- (c) Find  $E(Y|X = x)$  for  $x > 0$ .
- (d) Find  $P(X \leq 2, Y \leq 2)$ .
- (e) Find the correlation coefficient  $r(X, Y)$ .
- (f) Find another joint p.d.f. having the same marginals.

Ans. (b)  $f_1(x) = xe^{-x}$ ,  $0 < x < \infty$ ;  $f_2(y) = ye^{-y}$ ,  $0 < y < \infty$ .

(c)  $E(Y|x) = \frac{1-e^x}{x}[x-1] + \frac{1}{x}\left(\frac{x^2}{2} + xe^x + e^{-x} - 1\right)$

(d)  $1 - \frac{1}{e^4} - \frac{4}{e^2}$ ; (e)  $r(x, y) = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} = \frac{1}{2}$

(f) Hint.  $f(x, y, \alpha) = f_1(x)f_2(y)[1 + \alpha(2F(x)-1)(2F(y)-1)]$ ,  $|\alpha| < 1$ , has the same marginals  $f_1(x)$  and  $f_2(y)$ .

26. Obtain regression equation of  $Y$  on  $X$  for the distributions :

(a)  $f(x, y) = \frac{9}{2} \cdot \frac{1+x+y}{(1+x)^4(1+y)^4}$ ;  $x, y \geq 0$

(b)  $f(x, y) = \frac{4}{5}(x+3y)e^{-x-2y}$ ;  $x, y \geq 0$

[Sardar Patel Univ. M.Sc., 1992]

Ans. (a) Hint. See Example 5-25, page 5-55, (b)  $\frac{x+3}{2x+3}$ .

27. A ball is drawn at random from an urn containing three white balls numbered 0, 1, 2; two red balls numbered 0, 1 and one black ball numbered 0. If the colours white, red and black are again numbered 0, 1 and 2 respectively, find the correlation coefficient between the variates  $X$ , the colour number and  $Y$  the number of the ball. Write down the equation of regression line of  $Y$  on  $X$ .

[Calcutta Univ. B.Sc. (Maths. Hons.), 1986]

### OBJECTIVE TYPE QUESTIONS

I. State, giving reasons, whether each of the following statements is true or false.

- (i) Both regression lines of  $Y$  on  $X$  and of  $X$  on  $Y$  do not intersect at all.
- (ii) In a bivariate regression,  $b_{YX} = \frac{1}{5}$ ,  $b_{XY} = 10$
- (iii) The regression coefficient of  $Y$  on  $X$  is 3.2 and that of  $X$  on  $Y$  is 0.8.
- (iv) There is no relationship between correlation coefficient and regression coefficient.
- (v) Both the regression coefficients cannot exceed unity.

- (vi) The greater the value of ' $r$ ', the better are the estimates obtained through regression analysis.
- (vii) If  $X$  and  $Y$  are negatively correlated variables, and  $(0, 0)$  is on the least squares line of  $Y$  on  $X$ , and if  $X = 1$  is the observed value then predicted value of  $Y$  must be negative.
- (viii) Let the correlation between  $X$  and  $Y$  be perfect and positive. Suppose the points  $(3, 5)$  and  $(1, 4)$  are on the regression lines. With this knowledge it is possible to determine the least squares line exactly.
- (ix) If the lines of regression are  $Y = \frac{1}{4}X$  and  $X = \frac{1}{9}Y + 1$ , then  $\rho = \frac{1}{6}$  and  $E(X | Y = 0) = 1$ .
- (x) In a bivariate distribution,  $b_{YX} = 2.8$  and  $b_{XY} = 0.3$ .

**II. Fill in the blanks :**

- (i) The regression analysis measures ... between  $X$  and  $Y$ .
- (ii) Lines of regression are ... if  $r_{XY} = 0$  and they are ... if  $r_{XY} = \pm 1$ .
- (iii) If the regression coefficients of  $X$  on  $Y$  and  $Y$  on  $X$  are  $-0.4$  and  $-0.9$  respectively then the correlation coefficient between  $X$  and  $Y$  is ...
- (iv) If the two regression lines are  $X + 3Y - 5 = 0$  and  $4X + 3Y - 8 = 0$ , then the correlation coefficient between  $X$  and  $Y$  is ...
- (v) If one of the regression coefficients is ... unity, the other must be ... unity.
- (vi) The farther the two regression lines cut each other, the ... will be the degree of correlation.
- (vii) When one regression coefficient is positive, the other would be ...
- (viii) The sign of regression coefficient is ... as that of correlation coefficient.
- (ix) Correlation coefficient is the... between regression coefficients.
- (x) Arithmetic mean of regression coefficients is ... correlation coefficient.
- (xi) When the correlation coefficient is zero, the two regression lines are ... and when it is  $\pm 1$ , then the regression lines are ...

**III. Indicate the correct answer :**

- (i) The regression line of  $Y$  on  $X$  (a) minimises total of the squares of horizontal deviations, (b) total of the squares of the vertical deviations, (c) both vertical and horizontal deviations, (d) none of these.
- (ii) The regression coefficients are  $b_2$  and  $b_1$ . Then the correlation coefficient  $r$  is (a)  $b_1/b_2$ , (b)  $b_2/b_1$ , (c)  $b_1b_2$  (d)  $\pm \sqrt{b_1 b_2}$ .
- (iii) The farther the two regression lines cut each other (a) the greater will be the degree of correlation, (b) the lesser will be the degree of correlation, (c) does not matter.

- (iv) If one regression coefficient is greater than unity, then the other must be (a) greater than the first one, (b) equal to unity, (c) less than unity, (d) equal to zero.
- (v) When the correlation coefficient  $r = \pm 1$ , then the two regression lines (a) are perpendicular to each other; (b) coincide, (c) are parallel to each other, (d) do not exist.
- (vi) The two lines of regression are given as  $\bar{X} + 2\bar{Y} - 5 = 0$  and  $2\bar{X} + 3\bar{Y} = 8$ . Then the mean values of  $X$  and  $Y$  respectively are (a) 2, 1, (b) 1, 2, (c) 2, 5, (d) 2, 3.
- (vii) The tangent of the angle between two regression lines is given as 0.6 and the s.d. of  $Y$  is known to be twice that of  $X$ . Then the value of correlation coefficient between  $X$  and  $Y$  is (a)  $-\frac{1}{2}$ , (b)  $\frac{1}{2}$ , (c) 0.7, (d) 0.3.

**IV.**  $\sigma_x$  and  $\sigma_y$  are the standard deviations of two correlated variables  $X$  and  $Y$  respectively in a large sample, and  $r$  is the sample correlation coefficient.

- (i) State the "Standard Error of Estimate" for linear regression of  $Y$  on  $X$ .
- (ii) What is the standard error in estimating  $Y$  from  $X$  if  $r = 0$ ?
- (iii) By how much is this error reduced if  $r$  is increased to 0.30?
- (iv) How large must  $r$  be in order to reduce this standard error to one-half its value for  $r = 0$ ?
- (v) Give your interpretations for the cases  $r = 0$  and  $r = 1$ .

**V.** Explain why we have two lines of regression.

**10-8. Correlation Ratio.** As discussed earlier, when variables are linearly related, we have the regression lines of one variable on another variable and correlation coefficient can be computed to tell us about the extent of association between them. However, if the variables are not linearly related but some sort of curvilinear relationship exists between them, the use of  $r$  which is a measure of the degree to which the relation approaches a straight line "law" will be misleading. We might come across bivariate distributions where  $r$  may be very low or even zero but the regression may be strong, or even perfect. *Correlation ratio* ' $\eta$ ' is the appropriate measure of curvilinear relationship between the two variables. Just as  $r$  measures the concentration of points about the straight line of best fit,  $\eta$  measures the concentration of points about the curve of best fit. If regression is linear  $\eta = r$ , otherwise  $\eta > r$  (cf. Remark 2, § 10-8.1).

**10-8.1. Measure of Correlation Ratio.** In the previous articles we have assumed that there is a single observed value  $Y$  corresponding to the given value  $x_i$  of  $X$  but sometimes there are more than one such value of  $Y$ .

Suppose corresponding to the values  $x_i$ , ( $i = 1, 2, \dots, m$ ) of the variable  $X$ , the variable  $Y$  takes the values  $y_{ij}$  with respective frequencies  $f_{ij}$ ,  $j = 1, 2, \dots, n$ .

Though all the  $x$ 's in the  $i$ th vertical array have the same value, the  $y$ 's are different. A typical pair of values in the  $i$ th array is  $(x_i, y_{ij})$ , with frequency  $f_{ij}$ .

Thus the first suffix  $i$  indicates the vertical array while the second suffix  $j$  indicates the positions of  $y$  in that array. Let

$$\sum_{j=1}^n f_{ij} = n_i \quad \text{and} \quad \sum_{i=1}^m \sum_{j=1}^n f_{ij} = \sum_{i=1}^m \left( \sum_{j=1}^n f_{ij} \right) = \sum_{i=1}^m n_i = N, \quad (\text{say}).$$

If  $\bar{y}_i$  and  $\bar{y}$  denote the means of the  $i$ th array and the overall mean respectively, then

$$\bar{y}_i = \frac{\sum_{j=1}^n f_{ij} y_{ij}}{\sum_{j=1}^n f_{ij}} = \frac{\sum_{j=1}^n f_{ij} y_{ij}}{n_i} = \frac{T_i}{n_i} \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^m \sum_{j=1}^n f_{ij} y_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} = \frac{\sum_{i=1}^m n_i \bar{y}_i}{\sum_{i=1}^m n_i} = \frac{T}{N}$$

In other words  $\bar{y}$  is the weighted mean of all the array means, the weights being the array frequencies.

**Def.** The correlation ratio of  $Y$  on  $X$ , usually denoted by  $\eta_{yx}$  is given by

$$\eta_{yx}^2 = 1 - \frac{\sigma_{ey}^2}{\sigma_y^2} \quad \dots (10.21)$$

where  $\sigma_{ey}^2$  and  $\sigma_y^2$  are given by

$$\sigma_{ey}^2 = \frac{1}{N} \sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 \quad \text{and} \quad \sigma_y^2 = \frac{1}{N} \sum_i \sum_j f_{ij} (y_{ij} - \bar{y})^2$$

A convenient expression for  $\eta_{yx}$  can be obtained in terms of standard deviation  $\sigma_{my}$  of the means of the vertical arrays, each mean being weighted by the array frequency.

We have

$$\begin{aligned} N\sigma_y^2 &= \sum_i \sum_j f_{ij} (y_{ij} - \bar{y})^2 = \sum_i \sum_j f_{ij} \{ (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) \}^2 \\ &= \sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 + \sum_i \sum_j f_{ij} (\bar{y}_i - \bar{y})^2 + 2 \sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) \end{aligned}$$

The term  $2[\sum_i (\bar{y}_i - \bar{y}) \{\sum_j f_{ij} (y_{ij} - \bar{y}_i)\}]$  vanishes since  $\sum_j f_{ij} (y_{ij} - \bar{y}_i) = 0$ ,

being the algebraic sum of the deviations from mean.

$$\begin{aligned} \therefore N\sigma_y^2 &= \sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \bar{y})^2 \\ \Rightarrow N\sigma_y^2 &= N\sigma_{ey}^2 + N\sigma_{my}^2 \Rightarrow \sigma_y^2 = \sigma_{ey}^2 + \sigma_{my}^2 \\ \Rightarrow 1 - \frac{\sigma_{ey}^2}{\sigma_y^2} &= \frac{\sigma_{my}^2}{\sigma_y^2} \end{aligned}$$

which on comparison with (10.21) gives

$$\eta_{yx}^2 = \frac{\sigma_{my}^2}{\sigma_y^2} = \frac{\sum_i n_i (\bar{y}_i - \bar{y})^2}{\sum_i \sum_j f_{ij} (y_{ij} - \bar{y})^2} \quad \dots (10.22)$$

We have

$$\begin{aligned} N\sigma_{mY}^2 &= \sum_i n_i (\bar{y}_i - \bar{y})^2 = \sum_i n_i \bar{y}_i^2 - N \bar{y}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N} \\ \therefore \quad \eta_{YX}^2 &= \left[ \sum_i \left( \frac{T_i^2}{n_i} \right) - \frac{T^2}{N} \right] / N\sigma_Y^2, \end{aligned} \quad \dots(10-23)$$

a formula, much more convenient for computational purposes.

**Remarks 1.** (10-21) implies that

$$\sigma_{eY}^2 = \sigma_Y^2 (1 - \eta_{YX}^2)$$

Since  $\sigma_{eY}^2$  and  $\sigma_Y^2$  are non-negative, we have

$$1 - \eta_{YX}^2 \geq 0 \Rightarrow \eta_{YX}^2 \leq 1 \Rightarrow |\eta_{YX}| \leq 1$$

2. Since the sum of squares of deviations in any array is minimum when measured from its mean, we have

$$\sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 \leq \sum_i \sum_j f_{ij} (y_{ij} - \hat{y}_{ij})^2 \quad \dots(*)$$

where  $\hat{y}_{ij}$  is the estimate of  $y_{ij}$  for given value of  $X = x_i$ , say, as given by the line of regression of  $Y$  on  $X$  i.e.,  $\hat{y}_{ij} = a + bx_i$ , ( $j = 1, 2, \dots, n$ ).

$$\text{But} \quad \sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 = N\sigma_{eY}^2 = N\sigma_Y^2 (1 - \eta_{YX}^2)$$

$$\text{and} \quad \sum_i \sum_j f_{ij} (y_{ij} - a - bx_i)^2 = N\sigma_Y^2 (1 - r^2) \quad (\text{c.f. } \S \text{ 10-7-6})$$

$$\therefore (*) \Rightarrow 1 - \eta_{YX}^2 \leq 1 - r^2$$

$$\text{i.e.,} \quad \eta_{YX}^2 \geq r^2 \Rightarrow |\eta_{YX}| \geq |r|$$

Thus the absolute value of the correlation ratio can never be less than the absolute of  $r$ , the correlation coefficient.

When the regression of  $Y$  on  $X$  is linear, straight line of means of arrays coincides with the line of regression and  $\eta_{YX}^2 = r^2$ . Thus  $\eta_{YX}^2 - r^2$  is the departure of regression from linearity. It is also clear (from Remark 1) that the more nearly  $\eta_{YX}^2$  approaches unity, the smaller is  $\sigma_{eY}^2$  and, therefore, closer are the points to the curve of means of vertical arrays.

$$\text{When} \quad \eta_{YX}^2 = 1, \sigma_{eY}^2 = 0 \Rightarrow \sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 = 0$$

$\Rightarrow y_{ij} = \bar{y}_i$ ,  $\forall j = 1, 2, \dots, n$ , i.e., all the points lie on the curve of means. This implies that there is a functional relationship between  $X$  and  $Y$ .  $\eta_{YX}$  is, therefore, the measure of the degree to which the association between the variables approaches a functional relationship of the form  $Y = F(X)$ , where  $F(X)$  is a single valued function of  $X$ ,  $[F(X) = a + bX]$ .

3. It is worth noting that the value of  $\eta_{YX}$  is not independent of the classification of the data. As the class intervals become narrower  $\eta_{YX}$  approaches unity, since in that case  $\sigma_{mY}^2$  gets nearer to  $\sigma_Y^2$ . If the grouping is so fine that only one item appears in each row (related to each  $x$ -class), that item will constitute the mean of that column and thus in this case  $\sigma_{mY}^2$  and  $\sigma_Y^2$  become equal so that  $\eta_{YX}^2 = 1$ . On the other hand, a very coarse grouping tends to make the value of  $\eta_{YX}$  approach  $r$ . "Student" has given a formula for 'the correction'

to be made in the correlation ratio 'for grouping' in Biometrika (Vol IX page 316-320.)

4. It can be easily proved that  $\eta_{YX}^2$  is independent of change of origin and scale of measurements.

5.  $\eta_{XY}^2$ , the second correlation ratio of  $X$  on  $Y$  depends upon the scatter of observations about the line of column means.

6.  $r_{XY}$  and  $r_{YX}$  are same but  $\eta_{YX}$  is, in general, different from  $\eta_{XY}$ .

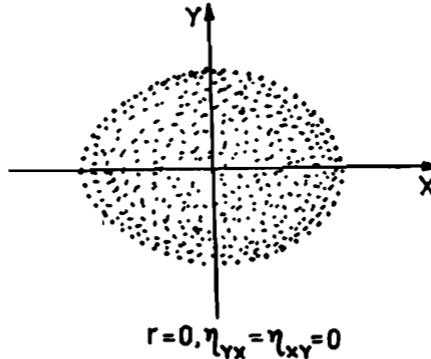
7. In terms of expectation, correlation ratio is defined as follows :

$$\eta_{YX}^2 = \frac{E_X [E(Y|X) - E(Y)]^2}{E[Y - E(Y)]^2} = \frac{E[E(Y|X) - E(Y)]^2}{\sigma_Y^2}$$

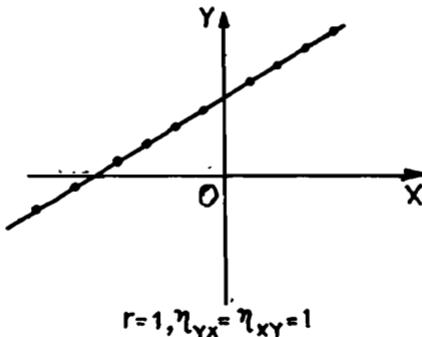
and  $\eta_{XY}^2 = \frac{E_Y [E(X|Y) - E(X)]^2}{E[X - E(X)]^2} = \frac{E[E(X|Y) - E(X)]^2}{\sigma_X^2}$

8. We give below some diagrams, exhibiting the relationship between  $r$  and  $\eta$ .

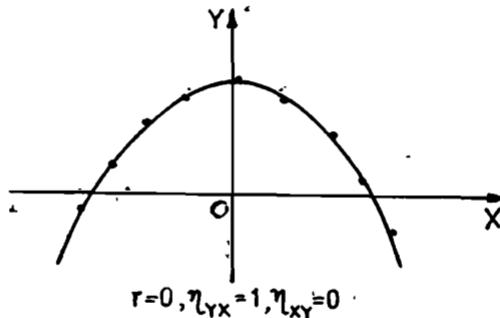
(i) For completely random scattering of the dots with no trend, both  $r$  and  $\eta$  are zero.



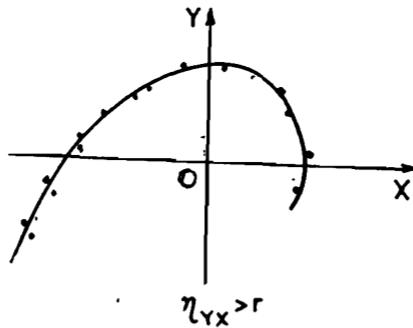
(ii) If dots lie precisely on a line,  $r = 1$  and  $\eta = 1$ .



- (iii) If dots lie on a curve, such that no ordinate cuts it more than once,  $\eta_{YX} = 1$  and if furthermore, the dots are symmetrically placed about  $Y$ -axis, then  $\eta_{XY} = 0, r = 0$ .



- (iv) If  $\eta_{YX} > r$ , the dots are scattered around a definitely curved trend line.



### EXERCISE 10(e)

1. (a) Define correlation coefficient and correlation ratio. When is the latter a more suitable measure of correlation than the former ? Show that the correlation ratio is never less than the correlation coefficient. What do you infer if the two are equal ? Further, show that none of these can exceed one.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

(b) Show that  $1 \geq \eta_{YX}^2 \geq r_{YX}^2 \geq 0$

Interpret each of the following statements.

- (i)  $r = 0$ , (ii)  $r^2 = 1$ , (iii)  $\eta^2 = 1$ , (iv)  $\eta^2 = r^2$  and (v)  $\eta = 0$

- (c) When the correlation coefficient is equal to unity, show that the two correlation ratios are also equal to unity. Is the converse true ?

- (d) Define correlation ratio  $\eta_{XY}$  and prove that

$$1 \geq \eta^2_{XY} \geq r^2,$$

where  $r$  is the coefficient of correlation between  $X$  and  $Y$ . Show further that  $(\eta^2_{XY} - r^2)$  is a measure of non-linearity of regression.

**2. For the joint p.d.f.**

$$f(x, y) = \frac{1}{2}x^3 \exp[-x(y+1)], \quad y > 0, x > 0 \\ = 0 \quad \text{otherwise},$$

find :

- (i) Two lines of regression.
- (ii) The regression curves for the means.
- (iii)  $r(X, Y)$ .
- (iv)  $\eta^2_{YX}$  and  $\eta^2_{XY}$ .

[Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1987]

**Ans.** (i)  $y = -\frac{1}{6}x + 1$  ;  $x = -\frac{2}{3}y + \frac{10}{3}$

(ii)  $y = E(Y|x) = \frac{1}{x}$  ;  $x = E(X|y) = \frac{4}{1+y}$

(iii)  $r(X, Y) = -\frac{1}{3}$  (iv)  $\eta^2_{YX} = \frac{1}{3}$ ,  $\eta^2_{XY} = \frac{1}{5}$

**3. Compute  $r(X, Y)$  and  $\eta_{YX}$  for the following data :**

$X$ :	0.5 — 1.5	1.5 — 2.5	2.5 — 3.5	3.5 — 4.5	4.5 — 5.5
$f$ :	20	30	35	25	15
$\bar{y}_i$ :	11.3	12.7	14.7	16.5	19.1

$\text{Var}(Y) = 9.61$

**Ans.**  $\eta_{YX} = 0.77$ ,  $r = 0.85$

**4. Compute  $\eta_{XY}$  for the following table :**

		$X$					
			47	52	57	62	67
$Y$	57	4	4	2	...	...	
	62	4	8	8	1	...	
	67	...	7	12	1	4	
	72	...	3	1	8	5	
	77	...	...	3	5	6	

**10-9. Intra-class Correlation.** Intra-class correlation means within class correlation. It is distinguishable from product moment correlation in as much as here both the variables measure the same characteristics. Sometimes specially in biological and agricultural study, it is of interest to know how the members of a family or group are correlated among themselves with respect to some one of their common characteristic. For example, we may require the correlation between the heights of brothers of a family or between yields of plots of an experimental block. In such cases both the variables measure the same characteristic, e.g., height and height or weight and weight. There is

nothing to distinguish one from the other so that one may be treated as  $X$ -variable and the other as the  $Y$ -variable.

Suppose we have  $A_1, A_2, \dots, A_n$  families with  $k_1, k_2, \dots, k_n$  members, each of which may be represented as

	$x_{11}$	$x_{21} \dots$	$x_{i1} \dots$	$x_{n1}$
	$x_{12}$	$x_{22}$	$x_{i2}$	$x_{n2}$
	$x_{1j}$	$x_{2j} \dots$	$x_{ij} \dots$	$x_{nj}$
	$x_{1k_1}$	$x_{2k_2} \dots$	$x_{ik_i} \dots$	$x_{nk_n}$

and let  $x_{ij}$  ( $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, k_i$ ) denote the measurement on the  $j$ th member in the  $i$ th family.

We shall have  $k_i(k_i - 1)$  pairs for the  $i$ th family or group like  $(x_{ij}, x_{il})$ ,  $j \neq l$ . There will be  $\sum_{i=1}^n k_i(k_i - 1) = N$  pairs for all the  $n$  families or groups. If we prepare a correlation table there will be  $k_i(k_i - 1)$  entries for the  $i$ th group or family and  $\sum_i k_i(k_i - 1) = N$  entries for all the  $n$  families or groups. The table is symmetrical about the principal diagonal. Such a table is called an *intra-class correlation table* and the correlation is called *intra-class correlation*.

In the bivariate table  $x_{il}$  occurs  $(k_i - 1)$  times,  $x_{il}$  occurs  $(k_i - 1)$  times, ...,  $x_{ik_i}$  occurs  $(k_i - 1)$  times, i.e., from the  $i$ th family we have  $(k_i - 1) \sum_j x_{ij}$  and hence for all the  $n$  families we have  $\sum_i (k_i - 1) \sum_j x_{ij}$  as the marginal frequency, the table being symmetrical about principal diagonal.

$$\therefore \bar{x} = \bar{y} = \frac{1}{N} \left[ \sum_i (k_i - 1) \sum_j x_{ij} \right]$$

Similarly,

$$\sigma_x^2 = \sigma_y^2 = \frac{1}{N} \left[ \sum_i (k_i - 1) \sum_j (x_{ij} - \bar{x})^2 \right]$$

Further

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{N} \sum_i \left[ \sum_{j \neq l} (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right], j \neq l \\ &= \frac{1}{N} \sum_i \left[ \sum_{j=1}^{k_i} \sum_{l=1}^{k_i} (x_{ij} - \bar{x})(x_{il} - \bar{x}) - \sum_{j=1}^{k_i} (x_{ij} - \bar{x})^2 \right] \end{aligned}$$

If we write  $\bar{x}_i = \sum_j x_{ij} / k_i$ , then

$$\begin{aligned}\sum_i \left[ \sum_{j=1}^k \sum_{l=1}^k (x_{ij} - \bar{x}) (x_{il} - \bar{x}) \right] &= \sum_i \left[ \sum_j (x_{ij} - \bar{x}) \sum_l (x_{il} - \bar{x}) \right] \\ &= \sum_i [k_i (\bar{x}_i - \bar{x}) k_i (\bar{x}_i - \bar{x})] \\ &= \sum_i k_i^2 (\bar{x}_i - \bar{x})^2\end{aligned}$$

Therefore intra-class correlation coefficient is given by

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X) V(Y)}} = \frac{\sum_i k_i^2 (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{\sum_i \sum_j (k_i - 1) (x_{ij} - \bar{x})^2} \quad \dots(10-24)$$

If we put  $k_i = k$ , i.e., if all families have equal members then

$$\begin{aligned}r &= \frac{k^2 \sum_i (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{(k-1) \sum_i \sum_j (x_{ij} - \bar{x})^2} \\ &= \frac{nk^2 \sigma_m^2 - nk\sigma^2}{(k-1) nk\sigma^2} = \frac{1}{(k-1)} \left\{ \frac{k \sigma_m^2}{\sigma^2} - 1 \right\} \quad \dots(10-24a)\end{aligned}$$

where  $\sigma^2$  denotes the variance of  $X$  and  $\sigma_m^2$  the variance of means of families.

**Limits.** We have from (10-24a),

$$1 + (k-1)r = \frac{k\sigma_m^2}{\sigma^2} \geq 0 \Rightarrow r \geq -\frac{1}{(k-1)}$$

Also  $1 + (k-1)r \leq k$ , as the ratio  $\frac{\sigma_m^2}{\sigma^2} \leq 1 \Rightarrow r \leq 1$

so that  $-\frac{1}{(k-1)} \leq r \leq 1$

**Interpretation.** Intraclass correlation cannot be less than  $-1/(k-1)$ , though it may attain the value  $+1$  on the positive side, so that it is a skew coefficient and a negative value has not the same significance as a departure from independence as an equivalent positive value.

### EXERCISE 10 (f)

1. If  $x_1, x_2, \dots, x_k$  be  $k$  variates with standard deviation  $\sigma$  and  $m$  be any number, prove that

$$k^2 \sigma^2 = (k-1) \sum_{r=1}^k (x_r - m)^2 - \sum_{r=1}^k \sum_{s=1}^k (x_r - m)(x_s - m), \quad r \neq s$$

Hence deduce that the coefficient of intraclass correlation for  $n$  families with varying number of members in each family is

$$1 - \frac{\sum_i k_i \sigma_i^2}{\sigma^2 \sum_i k_i (k_i - 1)}$$

where  $k_i, \sigma_i^2$  denote the number of members and the variance respectively in the  $i$ th family and  $\sigma^2$  is the general variance.

Given  $n = 5, \sigma_i = i, k_i = i + 1 (i \leq 5)$ , find the least possible intraclass correlation coefficient.

2. What do you understand by intra-class correlation coefficient.

Calculate its value for the following data :

Family No.	<i>Height of brothers</i>			
1	60	62	63	65
2	59	60	61	62
3	62	62	64	63
4	65	66	65	66
5	66	67	67	69

$$h(y|x) dy = \frac{1}{\sigma_2 \sqrt{2\pi(1-\rho^2)}} \cdot \exp \left\{ -\frac{1}{2\sigma_2^2(1-\rho^2)} \left( y - \rho x \frac{\sigma_2}{\sigma_1} \right)^2 \right\}$$

The joint probability differential of  $X$  and  $Y$  is given by

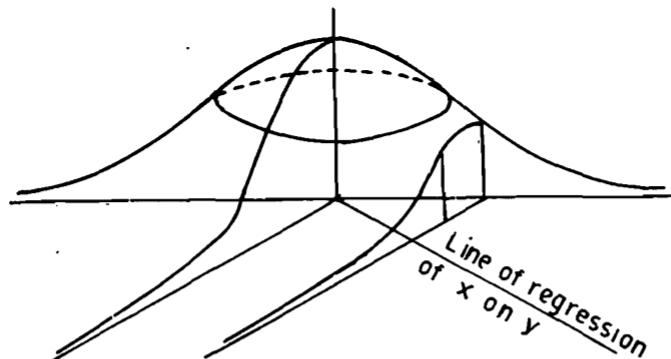
$$\begin{aligned} dP(x, y) &= g(x)h(y|x)dxdy \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot e^{-\frac{1}{2\sigma_1^2}x^2} e^{-\left[ -\frac{1}{2\sigma_2^2(1-\rho^2)} \left( y - \rho x \frac{\sigma_2}{\sigma_1} \right)^2 \right]} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot \exp \left\{ -\frac{1}{2(1-\rho^2)} \left( \frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right) \right\} \end{aligned}$$

Shifting the origin to  $(\mu_1, \mu_2)$ , we get

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot e^{-\frac{1}{2(1-\rho^2)} \left\{ \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right\}}; \quad (-\infty < x < \infty, -\infty < y < \infty) \quad \dots(10-25)$$

where  $\mu_1, \mu_2, \sigma_1 (>0), \sigma_2 (>0)$  and  $\rho (-1 < \rho < 1)$  are the five parameters of the distribution.

#### NORMAL CORRELATION SURFACE



This is the density function of a bivariate normal distribution. The variables  $X$  and  $Y$  are said to be normally correlated and the surface  $z = f(x, y)$  is known as the *normal correlation surface*. The nature of the normal correlation surface is indicated in the above diagram.

**Remarks 1.** The vector  $(X, Y)'$  following the joint p.d.f.  $f(x, y)$  as given in (10-25), will be abbreviated as  $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$  or  $BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ . If in particular  $\mu_1 = \mu_2 = 0$  and  $\sigma_1 = \sigma_2 = 1$  then

$(X, Y) \sim N(0, 0, 1, 1, \rho)$  or  $BVN(0, 0, 1, 1, \rho)$ .

**2.** The curve  $z = f(x, y)$  which is the equation of a surface in three dimensions, is called the '*Normal Correlation Surface*'.

**10-10-1. Moment Generating Function of Bivariate Normal Distribution.** Let  $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ . By def.,

$$\begin{aligned}
 M_{XY}(t_1, t_2) &= E[e^{t_1X + t_2Y}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(t_1x + t_2y) f(x, y) dx dy \\
 \text{Put } \frac{x - \mu_1}{\sigma_1} &= u, \frac{y - \mu_2}{\sigma_2} = v, -\infty < (u, v) < \infty \\
 \text{i.e., } x &= \sigma_1 u + \mu_1, y = \sigma_2 v + \mu_2 \Rightarrow |J| = \sigma_1 \sigma_2 \\
 \therefore M_{X,Y}(t_1, t_2) &= \frac{\exp(t_1\mu_1 + t_2\mu_2)}{2\pi\sqrt{1 - \rho^2}} \\
 &\times \iint_{u,v} \exp \left[ t_1\sigma_1 u + t_2\sigma_2 v - \frac{1}{2(1 - \rho^2)} \{ u^2 - 2\rho uv + v^2 \} \right] du dv \\
 &= \frac{\exp(t_1\mu_1 + t_2\mu_2)}{2\pi\sqrt{1 - \rho^2}} \\
 &\times \iint_{u,v} \exp \left[ \frac{1}{2(1 - \rho^2)} \{ (u^2 - 2\rho uv + v^2) - 2(1 - \rho^2)(t_1\sigma_1 u + t_2\sigma_2 v) \} \right] du dv
 \end{aligned}$$

We have

$$\begin{aligned}
 (u^2 - 2\rho uv + v^2) - 2(1 - \rho^2)(t_1\sigma_1 u + t_2\sigma_2 v) \\
 = [(u - \rho v) - (1 - \rho^2)t_1\sigma_1]^2 \\
 + (1 - \rho^2)\{(v - \rho t_1\sigma_1 - t_2\sigma_2)^2 - t_1^2\sigma_1^2 - t_2^2\sigma_2^2 - 2\rho t_1 t_2 \sigma_1 \sigma_2\} \quad \dots(*)
 \end{aligned}$$

By taking

$$\left. \begin{aligned}
 u - \rho v - (1 - \rho^2)t_1\sigma_1 &= \omega(1 - \rho^2)^{1/2} \\
 \text{and } v - \rho t_1\sigma_1 - t_2\sigma_2 &= z
 \end{aligned} \right\} \Rightarrow du dv = \sqrt{1 - \rho^2} dw dz$$

and using (\*), we get

$$\begin{aligned}
 M_{X,Y}(t_1, t_2) &= \exp[t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2\rho t_1 t_2 \sigma_1 \sigma_2)] \\
 &\times \left[ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\omega^2/2} d\omega \right] \times \left[ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz \right] \\
 &= \exp[t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2\rho t_1 t_2 \sigma_1 \sigma_2)] \quad \dots(10-26)
 \end{aligned}$$

In particular if  $(X, Y) \sim BVN(0, 0, 1, 1, \rho)$ , then

$$M_{X,Y}(t_1, t_2) = \exp[\frac{1}{2}(t_1^2 + t_2^2 + 2\rho t_1 t_2)] \quad \dots(10-26a)$$

**Theorem 10-5.** Let  $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ . Then X and Y are independent if and only if  $\rho = 0$ .

**Proof.** (a) If  $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$  and  $\rho = 0$ , then X and Y are independent [c.f. Remark 2(a) to Theorem 10-2, page 10-5].

Aliter.  $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

$$\therefore M_{X,Y}(t_1, t_2) = \exp(t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + 2\rho t_1 t_2 \sigma_1 \sigma_2 + t_2^2\sigma_2^2))$$

If  $\rho = 0$ , then

$$M_{X,Y}(t_1, t_2) = \exp(t_1\mu_1 + \frac{1}{2}t_1^2\sigma_1^2) \cdot \exp(t_2\mu_2 + \frac{1}{2}t_2^2\sigma_2^2)$$

$$\therefore = M_X(t_1) \cdot M_Y(t_2). \quad \dots(*)$$

[∴ If  $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , then the marginal p.d.f.'s of  $X$  and  $Y$  are normal i.e.,  $X \sim N(\mu_1, \sigma_1^2)$  and  $Y \sim N(\mu_2, \sigma_2^2)$ .]

(\*)  $\Rightarrow X$  and  $Y$  are independent.

(b) Conversely if  $X$  and  $Y$  are independent, then  $\rho = 0$  [c.f. Theorem 10-2]

**Theorem 10-6.**  $(X, Y)$  possesses a bivariate normal distribution if and only if every linear combination of  $X$  and  $Y$  viz.,  $aX + bY, a \neq 0, b \neq 0$ , is a normal variate.

**Proof.** (a) Let  $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , then we shall prove that  $aX + bY, a \neq 0, b \neq 0$  is a normal variate.

Since  $(X, Y)$  has a bivariate normal distribution, we have

$$M_{X,Y}(t_1, t_2) = E(e^{t_1X + t_2Y}) \\ = e^{t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + 2\rho t_1 t_2 \sigma_1 \sigma_2 + t_2^2\sigma_2^2)} \quad \dots(*)$$

Then m.g.f. of  $Z = aX + bY$ , is given by :

$$M_Z(t) = E(e^{tZ}) = E(e^{t(aX + bY)}) = E(e^{atX + btY}) \\ = \exp\{t(a\mu_1 + b\mu_2) + \frac{t^2}{2}(a^2\sigma_1^2 + 2ab\sigma_1\sigma_2 + b^2\sigma_2^2)\},$$

[Taking  $t_1 = at, t_2 = bt$  in (\*)]

which is the m.g.f. of normal distribution with parameters

$$\mu = a\mu_1 + b\mu_2, \sigma^2 = a^2\sigma_1^2 + 2ab\sigma_1\sigma_2 + b^2\sigma_2^2. \quad \dots(**)$$

Hence by uniqueness theorem of m.g.f.,

$$Z = aX + bY \sim N(\mu, \sigma^2),$$

where  $\mu$  and  $\sigma^2$  are given in (\*\*).

(b) Conversely, let  $Z = aX + bY, a \neq 0, b \neq 0$  be a normal variate. Then we have to prove that  $(X, Y)$  has a bivariate normal distribution.

Let  $Z = aX + bY \sim N(\mu, \sigma^2)$ ,

where  $\mu = EZ = E(aX + bY) = a\mu_x + b\mu_y$

and  $\sigma^2 = \text{Var } Z = \text{Var}(aX + bY) = a^2\sigma_x^2 + 2ab\rho\sigma_x\sigma_y + b^2\sigma_y^2$

$$\therefore M_Z(t) = \exp[t\mu + t^2\sigma^2/2] \\ = \exp[t(a\mu_x + b\mu_y) + \frac{t^2}{2}(a^2\sigma_x^2 + 2ab\rho\sigma_x\sigma_y + b^2\sigma_y^2)] \\ = \exp[t_1\mu_x + t_2\mu_y + \frac{1}{2}(t_1^2\sigma_x^2 + 2\rho t_1 t_2 \sigma_x \sigma_y + t_2^2\sigma_y^2)] \quad \dots(***)$$

where  $t_1 = at$  and  $t_2 = bt$ .

But (\*\*\* ) is the m.g.f. of  $BVN$  distribution with parameters  $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ . Hence by uniqueness theorem of m.g.f.

$$(X, Y) \sim BVN(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$$

### 10.10.2. Marginal Distribution of Bivariate Normal Distribution.

The marginal distribution of random variable  $X$  is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

Put  $\frac{y - \mu_2}{\sigma_2} = u$ , then  $dy = \sigma_2 du$ . Therefore,

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\times \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)} \left\{ \left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho u \left(\frac{x-\mu_1}{\sigma_1}\right) + u^2 \right\}\right] \sigma_2 du \\ &= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2} \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \\ &\times \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)} \left\{ u - \rho \left(\frac{x-\mu_1}{\sigma_1}\right)\right\}^2\right] du \end{aligned}$$

Put  $\frac{1}{\sqrt{1-\rho^2}} \left[ u - \rho \left(\frac{x-\mu_1}{\sigma_1}\right) \right] = t$ , then  $du = \sqrt{1-\rho^2} dt$

$$\begin{aligned} \therefore f_X(x) &= \frac{2}{2\pi\sigma_1} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt \\ &= \frac{1}{2\pi\sigma_1} \exp\left[-\frac{1}{2} \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \sqrt{2\pi} \\ &= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \quad \dots(10.27) \end{aligned}$$

Similarly, we shall get

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x, y) dx \\ &= \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right] \quad (10.27a) \end{aligned}$$

$$\text{Hence } X \sim N(\mu_1, \sigma_1^2) \text{ and } Y \sim N(\mu_2, \sigma_2^2) \quad \dots(10.27b)$$

**Remark.** We have proved that if  $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , then the marginal p.d.f.'s of  $X$  and  $Y$  are also normal. However, the converse is not true, i.e., we may have joint p.d.f.  $f(X, Y)$  of  $(X, Y)$  which is not

normal but the marginal p.d.f.'s may still be normal as discussed in the following illustration.

Consider the joint distribution of  $X$  and  $Y$  given by :

$$\begin{aligned} f(x, y) &= \frac{1}{2} \left[ \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp \left\{ \frac{-1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2) \right\} \right. \\ &\quad \left. + \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} (x^2 + 2\rho xy + y^2) \right\} \right] \\ &= \frac{1}{2} [f_1(x, y) + f_2(x, y)] ; -\infty < x, y < \infty \end{aligned} \quad \dots(10:27c)$$

where  $f_1(x, y)$  is the p.d.f. of  $BVN(0, 0, 1, 1, \rho)$  distribution and  $f_2(x, y)$  is the p.d.f. of  $BVN(0, 0, 1, 1, -\rho)$  distribution.

It can be easily verified that  $f(x, y)$  is the joint p.d.f. of  $(X, Y)$  and obviously  $f(x, y)$  is not the p.d.f. of bivariate normal distribution.

*Marginal distribution of  $X$*  in (10:27c)

$$f_X(x) = \frac{1}{2} \left[ \int_{-\infty}^{\infty} f_1(x, y) dy + \int_{-\infty}^{\infty} f_2(x, y) dy \right]$$

But  $\int_{-\infty}^{\infty} f_1(x, y) dy$  is the marginal p.d.f. of  $X$ , where

$(X, Y) \sim BVN(0, 0, 1, 1, \rho)$  and is given by  $X \sim N(0, 1)$ .

Similarly  $\int_{-\infty}^{\infty} f_2(x, y) dy$  is the marginal p.d.f. of  $X$ , where

$(X, Y) \sim BVN(0, 0, 1, 1, -\rho)$  and is given by  $X \sim N(0, 1)$ .

$$\begin{aligned} \therefore f_X(x) &= \frac{1}{2} \left[ \frac{1}{\sqrt{2\pi}} e^{-x^2/2} + \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \right] \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} ; -\infty < x < \infty \end{aligned} \quad \dots(i)$$

$\Rightarrow X \sim N(0, 1)$  i.e., the marginal distribution of  $X$  ... (10:27c) is normal.

Similarly, we can show that the marginal p.d.f. of  $Y$  in (10:27c) is given by :

$$\begin{aligned} f_Y(y) &= \frac{1}{2} \left[ \frac{1}{\sqrt{2\pi}} e^{-y^2/2} + \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \right] \\ &= \frac{1}{\sqrt{2\pi}} e^{-y^2/2} ; -\infty < y < \infty \end{aligned} \quad \dots(ii)$$

$\Rightarrow Y \sim N(0, 1)$ .

Hence if the marginal distributions of  $X$  and  $Y$  are normal (Gaussian), it does not necessarily imply that the joint distribution of  $(X, Y)$  is bivariate normal.

For another illustration, see Question Number 17, Exercise 10(f).

We further note that for the joint p.d.f. (10:27c), on using (i) and (ii), we have

$$E(X) = 0, \sigma_X^2 = 1 \text{ and } E(Y) = 0, \sigma_Y^2 = 1.$$

$$\therefore \text{Cov}(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} = E(XY)$$

$$= \frac{1}{2} \left[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_1(x, y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_2(x, y) dx dy \right]$$

$$= \frac{1}{2} [\rho + (-\rho)] = 0,$$

because, for  $f_1(x, y)$ ,  $(X, Y) \sim \text{BVN}(0, 0, 1, 1, \rho)$  and for  $f_2(x, y)$ ,  $(X, Y) \sim \text{BVN}(0, 0, 1, 1, -\rho)$ .

$$\therefore \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

However, we have : [From (i) and (ii)]

$$f_X(x) \cdot f_Y(y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2 + y^2)} \neq f(x, y)$$

$\Rightarrow X$  and  $Y$  are not independent.

The above example illustrates that we may have a joint density (non-Gaussian) of  $rv$ 's  $(X, Y)$  in which the marginal p.d.f.'s of  $X$  and  $Y$  are normal and  $\rho(X, Y) = 0$  and yet  $X$  and  $Y$  are not independent.

**10-10-3: Conditional Distributions.** Conditional distribution of  $X$  for a fixed  $Y$  is given by

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{XY}(x, y)}{f_Y(y)} \\ &= \frac{1}{\sqrt{2\pi} \sigma_1 \sqrt{1-\rho^2}} \exp \left[ -\frac{1}{2(1-\rho^2)} \left\{ \left( \frac{x-\mu_1}{\sigma_1} \right)^2 \right. \right. \\ &\quad \left. \left. - 2\rho \left( \frac{x-\mu_1}{\sigma_1} \right) \left( \frac{y-\mu_2}{\sigma_2} \right) + \left( \frac{y-\mu_2}{\sigma_2} \right)^2 \left( 1 - \frac{1}{1-\rho^2} \right) \right\} \right] \\ &= \frac{1}{\sigma_1 \sqrt{2\pi} \sqrt{1-\rho^2}} \\ &\quad \times \exp \left[ -\frac{1}{2(1-\rho^2)\sigma_1^2} \left\{ (x-\mu_1) - \rho \frac{\sigma_1}{\sigma_2} (y-\mu_2) \right\}^2 \right] \\ &= \frac{1}{\sqrt{2\pi} \sigma_1 \sqrt{1-\rho^2}} \\ &\quad \times \exp \left[ -\frac{1}{2(1-\rho^2)\sigma_1^2} \left\{ x - \left( \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y-\mu_2) \right) \right\}^2 \right] \end{aligned}$$

which is the probability function of a univariate normal distribution with mean and variance given by

$$E(X|Y=y) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2) \quad \text{and} \quad V(X|Y=y) = \sigma_1^2 (1 - \rho^2)$$

Hence the conditional distribution of  $X$  for fixed  $Y$  is given by :

$$(X \mid Y = y) \sim N \left[ \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2), \sigma_1^2 (1 - \rho^2) \right] \quad \dots(10.27d)$$

Similarly the conditional distribution of random variables  $Y$  for a fixed  $X$  is

$$\begin{aligned} f_{Y \mid X}(y \mid x) &= \frac{f_{XY}(x, y)}{f_X(x)} \\ &= \frac{1}{\sqrt{2\pi} \sigma_2 \sqrt{1 - \rho^2}} \\ &\times \exp \left[ -\frac{1}{2(1 - \rho^2) \sigma_2^2} \left\{ (y - \mu_2) - \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1) \right\}^2 \right], \\ &- \infty < y < \infty \end{aligned}$$

Thus the conditional distribution of  $Y$  for fixed  $X$  is given by

$$(Y \mid X = x) \sim N \left[ \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1), \sigma_2^2 (1 - \rho^2) \right] \quad \dots(10.27e)$$

It is apparent from the above results that the array means are collinear, i.e., the regression equations are linear (involving linear functions of the independent variables) and the array variances are constant (i.e., free from independent variable). We express this by saying that *the regression equations of  $Y$  on  $X$  and  $X$  on  $Y$  are linear and homoscedastic*.

For  $\rho = 0$ , the conditional variance  $V(Y \mid X)$  is equal to the marginal variance  $\sigma_2^2$  and the conditional mean  $E(Y \mid X)$  is equal to the marginal mean  $\mu_2$  and the two variables become independent, which is also apparent from joint distribution function. In between the two extremes when  $\rho = \pm 1$ , the correlation coefficient  $\rho$  provides a measure of degree of association or interdependence between the two variables.

**Example 10.27.** Show that for the bivariate normal distribution :

$$dP = \text{const. } \exp \left[ -\frac{1}{2(1 - \rho^2)} (x^2 - 2\rho xy + y^2) \right] dx dy,$$

$$(i) M.G.F. is M(t<sub>1</sub>, t<sub>2</sub>) = exp. \left[ \frac{1}{2} (t_1^2 + 2\rho t_1 t_2 + t_2^2) \right]$$

(ii) Moments obey the recurrence relation,

$$\mu_{rs} = (r + s - 1) \rho \mu_{r-1, s-1} + (r - 1)(s - 1)(1 - \rho^2) \mu_{r-2, s-2}$$

Hence or otherwise, show that

$$\mu_{rs} = 0, \text{ if } r + s \text{ is odd}, \mu_{31} = 3\rho, \mu_{22} = 1 + 2\rho^2$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

**Solution.** (i) From the given probability function, we see that

$$\mu_1 = 0 = \mu_2 \text{ and } \sigma_1^2 = 1 = \sigma_2^2.$$

∴ From (10.26a), we get

$$M = M(t_1, t_2) = \exp \left[ \frac{1}{2} (t_1^2 + 2\rho t_1 t_2 + t_2^2) \right]$$

$$(ii) \quad \frac{\partial M}{\partial t_1} = M(t_1 + \rho t_2) \text{ and } \frac{\partial M}{\partial t_2} = M(t_2 + \rho t_1)$$

and  $\frac{\partial^2 M}{\partial t_1 \partial t_2} = \frac{\partial}{\partial t_1} \left( \frac{\partial M}{\partial t_2} \right) = \frac{\partial}{\partial t_1} [M(t_2 + \rho t_1)]$   
 $= M\rho + (t_2 + \rho t_1)(t_1 + \rho t_2)M$   
 $\therefore \frac{\partial^2 M}{\partial t_1 \partial t_2} - \rho t_1 \frac{\partial M}{\partial t_1} - \rho t_2 \frac{\partial M}{\partial t_2}$   
 $= [M\rho + (t_2 + \rho t_1)(t_1 + \rho t_2)M] - \rho t_1 (t_1 + \rho t_2)M - \rho t_2 (t_2 + \rho t_1)M$   
 $= M[t_1 t_2 + \rho - \rho^2 t_1 t_2] \quad (\text{On simplification})$   
 $= M\rho + (1 - \rho^2)M t_1 t_2$   
 $\therefore \frac{\partial^2 M}{\partial t_1 \partial t_2} = \rho t_1 \frac{\partial M}{\partial t_1} + \rho t_2 \frac{\partial M}{\partial t_2} + M\rho + M(1 - \rho^2)t_1 t_2 \quad \dots (*)$

But  $M = \exp \left[ \frac{1}{2} (t_1^2 + 2\rho t_1 t_2 + t_2^2) \right] = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \mu_{rs} \cdot \frac{t_1^r t_2^s}{r! s!}$

$\therefore (*)$  gives

$$\begin{aligned} & \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} \mu_{rs} \cdot \frac{t_1^{r-1} t_2^{s-1}}{(r-1)! (s-1)!} \\ &= \left[ \rho \sum_{r=1}^{\infty} \sum_{s=0}^{\infty} r \mu_{rs} \cdot \frac{t_1^r t_2^s}{r! s!} + \rho \sum_{r=0}^{\infty} \sum_{s=1}^{\infty} s \mu_{rs} \cdot \frac{t_1^r t_2^s}{r! s!} \right. \\ & \quad \left. + \rho \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \mu_{rs} \cdot \frac{t_1^r t_2^s}{r! s!} + (1 - \rho^2) \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \mu_{rs} \cdot \frac{t_1^{r+1} t_2^{s+1}}{r! s!} \right] \end{aligned}$$

Equating the coefficients of  $\frac{t_1^{r-1}}{(r-1)!} \cdot \frac{t_2^{s-1}}{(s-1)!}$  on both sides, we get

$$\mu_{rs} = [\rho(r-1) \mu_{r-1,s-1} + \rho(s-1) \mu_{r-1,s-1} + \rho^2 \mu_{r-1,s-1} + (1 - \rho^2)(r-1)(s-1) \mu_{r-2,s-2}]$$

$$\Rightarrow \mu_{rs} = (r+s-1) \rho \mu_{r-1,s-1} + (r-1)(s-1)(1-\rho^2) \mu_{r-2,s-2}$$

In particular

$$\mu_{31} = 3\rho \mu_{2,0} + 0 = 3\rho \sigma_1^2 = 3\rho \quad (\because \sigma_1^2 = 1)$$

$$\mu_{22} = 3\rho \mu_{1,1} + (1 - \rho^2) \mu_{0,0} = 3\rho^2 + (1 - \rho^2).1$$

$$= (1 + 2\rho^2) \quad (\because \mu_{11} = \rho \sigma_1 \sigma_2 = \rho)$$

Also  $\mu_{03} = \mu_{30} = 0$

$$\mu_{12} = 2\rho \mu_{0,1} + 0 = 0 \quad (\because \mu_{01} = \mu_{10} = 0)$$

$$\mu_{23} = 4\rho \mu_{1,2} + 1 \cdot 2 (1 - \rho^2) \mu_{0,1} = 0$$

Similarly, we will get  $\mu_{21} = 0, \mu_{32} = 0$

If  $r+s$  is odd, so is  $(r-1)+(s-1), (r-2)+(s-2)$ , and so on.

And since  $\mu_{30} = 0 = \mu_{03}, \mu_{12} = 0 = \mu_{21}, \mu_{23} = 0 = \mu_{32}, \dots$ , we finally get,

$$\mu_{rs} = 0, \text{ if } r+s \text{ is odd.}$$

**Example 10-28.** Show that if  $X_1$  and  $X_2$  are standard normal variates with correlation coefficient  $\rho$  between them, then the correlation coefficient between  $X_1^2$  and  $X_2^2$  is given by  $\rho^2$ .

**Solution.** Since  $X_1$  and  $X_2$ , are two standard normal variates, we have

$$E(X_1) = E(X_2) = 0 \text{ and } V(X_1) = E(X_1^2) = 1 = V(X_2) = E(X_2^2)$$

$$\therefore M_{X_1, X_2}(t_1, t_2) = \exp \left[ \frac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2) \right] \quad [\text{c.f. (10-26)}]$$

$$\text{Now } \rho(X_1^2, X_2^2) = \frac{E(X_1^2 X_2^2) - E(X_1^2) E(X_2^2)}{\sqrt{[E(X_1^4) - (E(X_1^2))^2]} \sqrt{[E(X_2^4) - (E(X_2^2))^2]}}$$

$$\text{where } E(X_1^2 X_2^2) = \text{Coefficient of } \frac{t_1^2 \cdot t_2^2}{2! \cdot 2!} \text{ in } M(t_1, t_2) = (2\rho^2 + 1)$$

$$E(X_1^4) = \text{Coefficient of } \frac{t_1^4}{4!} \text{ in } M(t_1, t_2) = 3$$

$$E(X_2^4) = \text{Coefficient of } \frac{t_2^4}{4!} \text{ in } M(t_1, t_2) = 3$$

$$\therefore \rho(X_1^2, X_2^2) = \frac{2\rho^2 + 1 - 1}{\sqrt{(3-1)} \sqrt{(3-1)}} = \rho^2$$

**Example 10-29.** The variables  $X$  and  $Y$  with zero means and standard deviations  $\sigma_1$  and  $\sigma_2$  are normally correlated with correlation coefficient  $\rho$ . Show that  $U$  and  $V$  defined as

$$U = \frac{X}{\sigma_1} + \frac{Y}{\sigma_2} \quad \text{and} \quad V = \frac{X}{\sigma_1} - \frac{Y}{\sigma_2}$$

are independent normal variates with variances  $2(1 + \rho)$  and  $2(1 - \rho)$  respectively.

**Solution.** We are given that

$$dF(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left[ -\frac{1}{2(1-\rho^2)} \left\{ \frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right\} \right] dx dy \quad -\infty < (x, y) < \infty$$

$$\text{Also } u = \frac{x}{\sigma_1} + \frac{y}{\sigma_2}, v = \frac{x}{\sigma_1} - \frac{y}{\sigma_2}$$

$$\therefore \frac{1}{2}(u+v) = \frac{x}{\sigma_1} \text{ and } \frac{1}{2}(u-v) = \frac{y}{\sigma_2} \Rightarrow x = \frac{\sigma_1}{2}(u+v) \text{ and } y = \frac{\sigma_2}{2}(u-v)$$

Jacobian of transformation  $J$  is given by

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{1}{2}\sigma_1 & \frac{1}{2}\sigma_1 \\ \frac{1}{2}\sigma_2 & -\frac{1}{2}\sigma_2 \end{vmatrix} = -\frac{\sigma_1\sigma_2}{2}$$

$$\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} = \frac{1}{4} [(u+v)^2 + (u-v)^2] = \frac{1}{2}(u^2 + v^2)$$

$$dF_1(u, v) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$\times \exp \left[ -\frac{1}{2(1-\rho^2)} \left\{ \frac{1}{2}(u^2 + v^2) - 2\rho \left( \frac{u^2 - v^2}{4} \right) \right\} \right] \frac{\sigma_1\sigma_2}{2} du dv$$

$$\begin{aligned}
 &= \frac{1}{2\pi \cdot 2\sqrt{(1-\rho^2)}} \exp \left[ -\frac{1}{4(1-\rho^2)} \{ (1-\rho)u^2 + (1+\rho)v^2 \} \right] du dv \\
 &= \frac{1}{2\pi \sqrt{2(1-\rho)} \sqrt{2(1+\rho)}} \exp \left[ -\frac{u^2}{2(1+\rho)2} - \frac{v^2}{2(1-\rho)2} \right] du dv \\
 &= \left[ \frac{1}{\sqrt{2\pi} \sqrt{2(1+\rho)}} \cdot \exp \left\{ -\frac{u^2}{2(1+\rho)2} \right\} \right] du \\
 &\quad \times \left[ \frac{1}{\sqrt{2\pi} \sqrt{2(1-\rho)}} \cdot \exp \left\{ -\frac{v^2}{2(1-\rho)2} \right\} \right] dv \\
 &= [f_1(u)du][f_2(v)dv], \text{ (say)}
 \end{aligned}$$

where  $f_1(u) = \frac{1}{\sqrt{2\pi} \sqrt{2(1+\rho)}} \cdot \exp \left\{ -\frac{u^2}{2(1+\rho)2} \right\}$

and  $f_2(v) = \frac{1}{\sqrt{2\pi} \sqrt{2(1-\rho)}} \cdot \exp \left\{ -\frac{v^2}{2(1-\rho)2} \right\}$

Hence  $U$  and  $V$  are independently distributed,  $U$  as  $N[0, 2(1+\rho)]$  and  $V$  as  $N[0, 2(1-\rho)]$ .

**Aliter.** Find joint m.g.f. of  $U$  and  $V$  viz.,

$$M(t_1, t_2) = E(e^{t_1 U + t_2 V}) = E[e^{X(t_1 + t_2)/\sigma_1 + Y(t_1 - t_2)/\sigma_2}]$$

and use  $E(e^{t_1 X + t_2 Y}) = \exp[(t_1^2 \sigma_1^2 + t_2^2 \sigma_2^2 + 2\rho t_1 t_2 \sigma_1 \sigma_2)/2]$

**Example 10-30.** If  $X$  and  $Y$  are standard normal variates with co-efficient of correlation  $\rho$ , show that

- (i) Regression of  $Y$  on  $X$  is linear.
- (ii)  $X+Y$  and  $X-Y$  are independently distributed.
- (iii)  $Q = \frac{X^2 - 2\rho XY + Y^2}{(1-\rho^2)}$  is distributed like a chi-square, i.e., as that of the sum of the squares of standard normal variates.

(Madras Univ. B.E., 1990)

**Solution.** (i) c.f. § 10-10-3.

(ii) Let  $u = x + y$  and  $v = x - y$

$$dF(x, y) = \frac{1}{2\pi \sqrt{1-\rho^2}} \exp \left[ -\frac{1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2) \right] dx dy$$

Now  $x = \frac{u+v}{2}$ ,  $y = \frac{u-v}{2}$

$$\therefore J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

$$dG(u, v) = C \exp \left[ -\frac{1}{2(1-\rho^2) \cdot 4} \{ 2(u^2 + v^2) - 2\rho(u^2 - v^2) \} \right] du dv$$

$$\text{where } C = \frac{1}{4\pi\sqrt{1-\rho^2}}$$

$$\begin{aligned} \therefore dG(u, v) &= C \exp \left[ -\frac{1}{4(1-\rho^2)} \{ (1-\rho)u^2 + (1+\rho)v^2 \} \right] du dv \\ &= \left[ C_1 \exp \left[ -\frac{u^2}{4(1+\rho)} \right] du \right] \times \left[ C_2 \exp \left[ -\frac{v^2}{4(1-\rho)} \right] dv \right] \\ &= [g_1(u)du] [g_2(v)dv], \text{ (say).} \end{aligned}$$

Hence  $U$  and  $V$  are independently distributed.

$$\begin{aligned} (iii) \quad M_Q(t) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{tQ} dF(x, y) \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(tQ) \\ &\quad \times \exp \left[ -\frac{1}{2(1-\rho^2)} \{ x^2 - 2\rho xy + y^2 \} \right] dx dy \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left( tQ - \frac{Q}{2} \right) dx dy \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left[ -\frac{Q}{2}(1-2t) \right] dx dy \end{aligned}$$

Put  $\sqrt{(1-2t)}x = u$  and  $\sqrt{(1-2t)}y = v$

$$\therefore dx = \frac{du}{\sqrt{(1-2t)}} \text{ and } dy = \frac{dv}{\sqrt{(1-2t)}}$$

$$\text{Also } Q = \frac{1}{(1-\rho^2)} [x^2 - 2\rho xy + y^2] = \frac{1}{(1-\rho^2)} \left[ \frac{u^2 - 2\rho uv + v^2}{1-2t} \right]$$

$$\begin{aligned} \therefore M_Q(t) &= \frac{1}{2\pi\sqrt{1-\rho^2}(1-2t)} \\ &\quad \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2(1-\rho^2)} (u^2 - 2\rho uv + v^2) \right] du dv \\ &= \frac{1}{(1-2t)} \cdot 1 = (1-2t)^{-1} \end{aligned}$$

which is the m.g.f. of chi-square ( $\chi^2$ ) variate\* with  $n (=2)$  degrees of freedom.

**Example 10-31.** Let  $X$  and  $Y$  be independent standard normal variates. Obtain the m.g.f. of  $XY$ . [Gauhati Univ. M.Sc., 1992]

**Solution.** We have, by definition :

$$M_{XY}(t) = E(e^{tXY}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{txy} \cdot f(x, y) dx dy$$

Since  $X$  and  $Y$  are independent standard normal variates, their joint p.d.f.  $f(x, y)$  is given by :

$$\begin{aligned} f(x, y) &= f_1(x) \cdot f_2(y) = \frac{1}{2\pi} e^{-x^2/2} e^{-y^2/2}; -\infty < (x, y) < \infty \\ \therefore M_{XY}(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2 - 2txy + y^2)} dx dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2(1-t^2)} \right. \\ &\quad \times \left. \left\{ \frac{x^2}{1/(1-t^2)} - \frac{2txy}{(1/\sqrt{1-t^2})(1/\sqrt{1-t^2})} + \frac{y^2}{1/(1-t^2)} \right\} \right] dx dy \end{aligned} \dots (*)$$

If  $(U, V) \sim BVN(0, 0, \sigma_1^2, \sigma_2^2, \rho)$ , then we have

$$\begin{aligned} &\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)} \left\{ \frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right\}} dx dy = 1 \\ &\Rightarrow \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)} \left\{ \frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right\}} dx dy = 2\pi\sigma_1\sigma_2\sqrt{1-\rho^2} \quad \dots (**) \end{aligned}$$

Comparing (\*) and (\*\*) with

$$\sigma_1^2 = \sigma_2^2 = \frac{1}{(1-t^2)} \text{ and } \rho = t, \text{ we get}$$

$$\begin{aligned} M_{XY}(t) &= \frac{1}{2\pi} \cdot 2\pi \frac{1}{\sqrt{1-t^2}} \cdot \frac{1}{\sqrt{1-t^2}} \cdot \sqrt{1-t^2} \\ \Rightarrow M_{XY}(t) &= (1-t^2)^{1/2}; -1 < t < 1 \end{aligned}$$

**Example 10-32.** Let  $X$  and  $Y$  have bivariate normal distribution with parameters :

$$\mu_X = 5, \mu_Y = 10, \sigma_X^2 = 1, \sigma_Y^2 = 25 \text{ and } \text{Corr}(X, Y) = \rho.$$

(a) If  $\rho > 0$ , find  $\rho$  when  $P(4 < Y < 16 | X = 5) = 0.954$

[Delhi Univ. B.Sc. (Math. Hons.), 1993, '83]

\*Chi-square distribution is discussed in Chapter 13

(b) If  $\rho = 0$ , find  $P(X + Y \leq 16)$ .

**Solution.** Since  $(X, Y) \sim BVN(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ , the conditional distribution of  $Y$  given  $X = x$  is also normal.

$$\begin{aligned} (Y|X=x) &\sim N\left[\mu = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X), \sigma^2 = \sigma_Y^2(1 - \rho^2)\right] \\ \therefore (Y|X=5) &\sim N\left[\mu = 10 + \rho \times \frac{5}{1}(5 - 5), \sigma^2 = 25(1 - \rho^2)\right] \\ &= N\left[\mu = 10, \sigma^2 = 25(1 - \rho^2)\right] \end{aligned}$$

We want  $\rho$  so that

$$P(4 < Y < 16 | X = 5) = 0.954$$

$$\text{where } Z = \frac{Y - \mu}{\sigma} = \frac{Y - 10}{5\sqrt{1 - \rho^2}} \sim N(0, 1)$$

$$\Rightarrow P\left(\frac{4 - 10}{\sigma} < Z < \frac{16 - 10}{\sigma}\right) = 0.954$$

$$\Rightarrow P\left(\frac{-6}{\sigma} < Z < \frac{6}{\sigma}\right) = 0.954 \quad \dots(*)$$

But we know that if  $Z \sim N(0, 1)$ , then

$$P(-2 < Z < 2) = 0.954 \quad \dots(**)$$

Comparing (\*) and (\*\*), we get

$$\frac{6}{\sigma} = 2 \Rightarrow \sigma = 3 \Rightarrow \sigma^2 = 9 = 25(1 - \rho^2)$$

$$\therefore 1 - \rho^2 = \frac{9}{25} \Rightarrow \rho^2 = \frac{16}{25} \Rightarrow \rho = \frac{4}{5} = 0.8 \quad (\because \rho > 0)$$

(b) Since  $(X, Y)$  have bivariate normal distribution,

$\rho = 0 \Rightarrow X$  and  $Y$  are independent rv's

and

$$X \sim N(\mu_X, \sigma_X^2) \text{ and } Y \sim N(\mu_Y, \sigma_Y^2)$$

$$\therefore X + Y \sim N(\mu = \mu_X + \mu_Y, \sigma^2 = \sigma_X^2 + \sigma_Y^2) = N(15, 26)$$

Hence

$$P(X + Y \leq 16) = P\left(Z \leq \frac{16 - 15}{\sqrt{26}}\right)$$

$$\text{where } Z = \frac{(X + Y) - \mu}{\sigma} \sim N(0, 1).$$

$$\therefore P(X + Y \leq 16) = P\left(Z \leq \frac{1}{\sqrt{26}}\right) = \Phi\left(\frac{1}{\sqrt{26}}\right),$$

where  $\Phi(z) = P(Z \leq z)$ , is the distribution function of standard normal variate.

$$\begin{aligned} \text{Remark . } P(X + Y \leq 16) &= P\left(Z \leq \frac{1}{5.099}\right) = P(Z \leq 0.196) \\ &= 0.5 + P(0 \leq Z \leq 0.196) \\ &= 0.5 + 0.0793 \text{ (approx.)} \\ &= 0.5793. \end{aligned}$$

## EXERCISE 10(f)

1. (a) Define conditional and marginal distributions. If  $X$  and  $Y$  follow bivariate normal distribution, find (i) the conditional distribution of  $X$  given  $Y$  and (ii) the marginal distribution of  $X$ . Show that the conditional mean of  $X$  is dependent on the given  $Y$ , but the conditional variance is independent of it.

(b) Define Bivariate Normal distribution. If  $(X, Y)$  has a bivariate normal distribution, find the marginal density function  $f_X(x)$  of  $X$ .

[Delhi Univ. B.Sc. (Maths. Hons.), 1988]

2. (a) The marks  $X$  and  $Y$  scored by candidates in an examination in two subjects Mathematics and Statistics are known to follow a bivariate normal distribution. The mean of  $X$  is 52 and its standard deviation is 15, while  $Y$  has mean 48 and standard deviation 13. Also the coefficient of correlation between  $X$  and  $Y$  is 0.6.

Write down the joint distribution of  $X$  and  $Y$ . If 100 marks in the aggregate are needed for a pass in the examination, show how to calculate the proportion of candidates who pass the examination?

(b) A manufacturer of electric bulbs, in his desire for putting only good bulbs for sale, rejects all bulbs for which a certain quality characteristic  $X$  of the filament is less than 65 units. Assume that the quality characteristic  $X$  and the life  $Y$ , of the bulb in hours are jointly normally distributed with parameters given below :

	$X$	$Y$
Mean	80	1100
Standard deviation	10	10

Correlation coefficient  $\rho(X, Y) = 0.60$

Find (i) the proportion of bulbs produced that will burn for less than 1000 hours, (ii) the proportion of bulbs produced that will be put for sale, (iii) the average life of bulbs put for sale.

3. (a) Determine the parameters of the bivariate normal distribution :

$$f(x, y) = k \exp \left[ -\frac{8}{27} \{ (x - 7)^2 - 2(x - 7)(y + 5) + 4(y + 5)^2 \} \right]$$

Also find the value of  $k$ .

(b) For the bivariate normal distribution :

$$(X, Y) \sim BVN \left( 1, 2, 4^2, 5^2, \frac{12}{13} \right)$$

find (i)  $P(X > 2)$ , (ii)  $P(X > 2 | Y = 2)$ .

(c) The bivariate random variable  $(X_1, X_2)$  have a bivariate normal distribution with means 60 and 75 and standard deviations 6 and 12 with a correlation coefficient of 0.55. Find the following probabilities :

(i)  $P(65 \leq X_1 \leq 75)$ , (ii)  $P(71 \leq X_2 \leq 80 | X_1 = 55)$  and (iii)  $P(|X_1 - X_2| \geq 15)$ .

4. For a bivariate normal distribution :

$$f_{XY}(x, y) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp \left\{ -\frac{1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2) \right\},$$

$-\infty < (x, y) < \infty$

- Find (i) marginal distribution of  $X$  and  $Y$ ,  
(ii) conditional distribution of  $Y$  given  $X$ ,  
(iii) distribution of  $\frac{1}{(1-\rho^2)} [x^2 - 2\rho xy + y^2]$ ,

and (iv) show that in general  $X$  and  $Y$  are stochastically dependent and will be independent if and only if  $\rho = 0$ .

5. Let the joint p.d.f. of  $X$  and  $Y$  be

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)}{\sigma_1} \frac{(y-\mu_2)}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}$$

where  $-\infty < x < \infty, -\infty < y < \infty, -1 < \rho < 1$ .

- (i) Find the marginal distribution of  $X$ .
- (ii) Find the conditional distribution of  $Y$  given  $X = x$ .
- (iii) Show that the regression of  $Y$  on  $X$  is linear and homoscedastic.
- (iv) Find  $P[3 < Y < 8 | X = 7]$ , given that  $\mu_1 = 3, \mu_2 = 1, \sigma_1^2 = 16, \sigma_2^2 = 25, \rho = 0.6$ ,
- (v) Find the probability of the simultaneous materialization of the inequalities,  $X > E(X)$  and  $Y > E(Y)$

**Hint.** (v) Required probability  $p$  is given by

$$p = P[X > E(X), Y > E(Y)] = P[X > \mu_1] \cap (Y > \mu_2)]$$

$$\begin{aligned} &= \int_{\mu_1}^{\infty} \int_{\mu_2}^{\infty} f(x, y) dx dy \\ &= \int_0^{\infty} \int_0^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left[ -\frac{1}{2(1-\rho^2)} (u^2 - 2\rho uv + v^2) \right] du dv, \\ &\quad \left( u = \frac{x-\mu_1}{\sigma_1}, v = \frac{y-\mu_2}{\sigma_2} \right). \end{aligned}$$

Now proceed as in Hint to Question Number 9(b).

6. Let the random variables  $X$  and  $Y$  be assumed to have a joint bivariate normal distribution with

$$\mu_1 = \mu_2 = 0, \sigma_1 = 4, \sigma_2 = 3 \text{ and } r(X, Y) = 0.8.$$

- (i) Write down the joint density function of  $X$  and  $Y$ .
- (ii) Write down the regression of  $Y$  on  $X$ .
- (iii) Obtain the joint density of  $X + Y$  and  $X - Y$ .

7. For the distribution of random variables  $X$  and  $Y$  given by

$$dF = k \exp \left[ -\frac{1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2) \right] dx dy, -\infty \leq x \leq \infty, -\infty \leq y \leq \infty$$

Obtain

- (i) the constant  $k$ ,
  - (ii) the distributions of  $X$  and  $Y$ ,
  - (iii) the distributions of  $X$  for given  $Y$  and of  $Y$  for given  $X$ ,
  - (iv) the curves of regression of  $\bar{Y}$  on  $X$  and of  $X$  on  $Y$ ,
- and (v) the distributions of  $X + Y$  and  $X - Y$ .

8. Let  $(X, Y)$  be a bivariate normal random variable with  $E(X) = E(Y) = 0$ ,  $\text{Var}(X) = \text{Var}(Y) = 1$  and  $\text{Cov}(X, Y) = \rho$ . Show that the random variable  $Z = Y/X$  has a Cauchy distribution.

[Delhi Univ. B.Sc. (Maths. Hons.), 1989]

$$\text{Ans. } f(z) = \frac{1}{\pi} \left[ \frac{(1-\rho^2)^{1/2}}{(1-\rho^2)+(z-\rho)^2} \right], -\infty < z < \infty.$$

9. (a) If  $(X, Y) \sim N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , prove that

$$P(X > \mu_x \cap Y > \mu_y) = \frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}$$

[Delhi Univ. M.Sc. (Stat.), 1987]

(b) If  $(X, Y) \sim N(0, 0, 1, 1, \rho)$  then prove that

$$P(X > 0 \cap Y > 0) = \frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}.$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

**Hint.**  $p = P(X > 0 \cap Y > 0)$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \times \int_0^\infty \int_0^\infty \exp \left[ -\frac{1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2) \right] dx dy$$

Put  $x = r \cos \theta$ ,  $y = r \sin \theta \Rightarrow |J| = r$ ;  $0 < r < \infty$ ,  $0 \leq \theta \leq \pi/2$

$$\therefore p = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_0^\infty \int_0^{\pi/2} \exp \left[ -\frac{r^2}{2(1-\rho^2)} (1 - \rho \sin 2\theta) \right] r dr d\theta$$

Now integrate first w.r. to  $r$  and then w.r. to  $\theta$ .

10. (a) Let  $X_1$  and  $X_2$  be two independent normally distributed variables with zero means and unit variances. Let  $Y_1$  and  $Y_2$  be the linear functions of  $X_1$  and  $X_2$  defined by

$$Y_1 = m_1 + l_{11}X_1 + l_{12}X_2, \quad Y_2 = m_2 + l_{21}X_1 + l_{22}X_2$$

Show that  $Y_1$  and  $Y_2$  are normally distributed with means  $m_1$  and  $m_2$ , variances  $\mu_{20} = l_{11}^2 + l_{12}^2$ ,  $\mu_{02} = l_{21}^2 + l_{22}^2$ , and covariance  $\mu_{12} = l_{11}l_{21} + l_{12}l_{22}$ .

(b) Let  $X_1$  and  $X_2$  be independent standard normal variates. Show that the variates  $Y_1, Y_2$  defined by

$Y_1 = a_1 + b_{11}X_1 + b_{12}X_2, \quad Y_2 = a_2 + b_{21}X_1 + b_{22}X_2$  are dependent normal variates and find their mean and variance.

**Hint.**  $Y_1$  and  $Y_2$ , being linear combination of S.N.V's are also normally distributed. To prove that they are dependent, it is sufficient to prove that  $r(Y_1, Y_2) \neq 0$ . [c.f. Remark 2 to Theorem 10.2]

11. (a) Show that, if  $X$  and  $Y$  are independent normal variates with zero means and variances  $\sigma_1^2$  and  $\sigma_2^2$  respectively, the point of inflexion of the curve of intersection of the normal correlation surface by planes through the  $z$ -axis, lie on the elliptical cylinder,

$$\frac{X^2}{\sigma_1^2} + \frac{Y^2}{\sigma_2^2} = 1$$

(b) If  $X$  and  $Y$  are bivariate normal variates with standard deviations unity and with correlation coefficient  $\rho$ , show that the regression of  $X^2$  ( $Y^2$ ) on  $Y^2$  ( $X^2$ ) is strictly linear. Also show that the regression of  $X$  ( $Y$ ) on  $Y^2$  ( $X^2$ ) is not linear.

12. For the bivariate normal distribution :

$$dF = k \exp \left[ -\frac{2}{3} (x^2 - xy + y^2 - 3x + 3y + 3) \right] dx dy,$$

obtain (i) the marginal distribution of  $Y$ , and

(ii) the conditional distribution of  $Y$  given  $X$ .

Also obtain the characteristic function of the above bivariate normal distribution and hence the covariance between  $X$  and  $Y$ .

13. Let  $f$  and  $g$  be the p.d.f.'s with corresponding distribution functions  $F$  and  $G$ . Also let

$$h(x, y) = f(x) g(y) [1 + \alpha (2F(x) - 1) (2G(y) - 1)],$$

where  $|\alpha| \leq 1$ , is a constant and  $h$  is a bivariate p.d.f. with marginal p.d.f.'s  $f$  and  $g$ . Further let  $f$  and  $g$  be p.d.f.'s of  $N(0, 1)$  distribution. Then prove that :

$$\text{Cov}(X, Y) = \alpha/\pi$$

14. If  $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , compute the correlation coefficient between  $e^X$  and  $e^Y$ .

Hint. Let  $U = e^X$ ,  $V = e^Y$ .

$$\begin{aligned}\mu'_{rs} &= E(U^r V^s) = E[e^{rX+sY}] \\ &= \exp[r\mu_1 + s\mu_2 + \frac{1}{2}(r^2\sigma_1^2 + s^2\sigma_2^2 + 2\rho rs)]\end{aligned}$$

[c.f. m.g.f. of B.V.N. distribution :  $t_1 = r$ ,  $t_2 = s$ ]

Now  $E(U) = \mu'_{10}$ ;  $E(U^2) = \mu'_{20}$ ,  $E(UV) = \mu_{11}'$  and so on.

$$\text{Ans. } \rho(U, V) = \frac{e^{\rho\sigma_1\sigma_2} - 1}{[(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)]^{1/2}}$$

15. If  $(X, Y) \sim BVN(0, 0, 1, 1, \rho)$ , find  $E[\max(X, Y)]$ .

$$\text{Hint. } \max(X, Y) = \frac{1}{2}(X + Y) + \frac{1}{2}|X - Y|$$

and  $Z = X - Y \sim N[0, 2(1 - \rho)]$  [c.f. Theorem 10.6]

$$\text{Ans. } E[\max(X, Y)] = \left( \frac{1 - \rho}{\pi} \right)^{1/2}$$

16. If  $(X, Y) \sim BVN(0, 0, 1, 1, \rho)$  with joint p.d.f.  $f(x, y)$  then prove that

$$(a) \quad P(XY > 0) = \frac{1}{2} + \frac{1}{\pi} \cdot \sin^{-1}(\rho).$$

**Hint.**  $P(XY > 0) = P(X > 0 \cap Y > 0) + P(X < 0 \cap Y < 0)$   
 $= 2 P(X > 0 \cap Y > 0)$  [By symmetry]

Now proceed as in Hint to Question No. 9(b).

$$(b) \quad 2\pi \int_{-\infty}^0 \int_{-\infty}^0 f(x, y) dx dy = \pi + \sin^{-1} \rho$$

17. The joint density of r.v's  $(X, Y)$  is given by :

$$f(x, y) = \frac{1}{2\pi} \cdot \exp[-(x^2 + y^2)/2] \times [1 + xy \exp[-(x^2 + y^2 - 2)/2]]; \\ -\infty < (x, y) < \infty$$

(i) Verify that  $f(x, y)$  is a p.d.f.

(ii) Show that the marginal distribution of each of  $X$  and  $Y$  is normal.

(iii) Are  $X$  and  $Y$  independent?

**Ans.** (ii)  $X \sim N(0, 1)$ ,  $Y \sim N(0, 1)$

(ii)  $X$  and  $Y$  are not independent.

18. Show by means of an example that the normality of conditional p.d.f.'s does not imply that the bivariate density is normal.

**Hint.** Consider  $f(x, y) = \text{constant. } \exp[-(1+x^2)(1+y^2)]$ ;  $-\infty < (x, y) < \infty$

. Then  $(Y|x) \sim N\left(0, \frac{1}{2(1+x^2)}\right)$  and  $(X|y) \sim N\left(0, \frac{1}{2(1+y^2)}\right)$

19. For a bivariate normal r.v.  $(X, Y)$ , does the conditional p.d.f. of  $(X, Y)$  given  $X + Y = c$ , (constant) exist? If so find it. If not, why not?

**Ans.** No, since  $P(X + Y = c) = 0$ .

20. Let

$$f(x, y) = \frac{1}{2} \left[ \begin{array}{l} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\} \\ + \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 + 2\rho xy + y^2)\right\} \end{array} \right] \\ -\infty < x < \infty, -\infty < y < \infty$$

then show that :

(i)  $f(x, y)$  is a joint p.d.f. such that both marginal densities are normal but  $f(x, y)$  is not bivariate normal.

(ii)  $X$  and  $Y$  have zero correlation but  $X$  and  $Y$  are not independent.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

21. Let  $X, Y$  be normally correlated variates with zero means and variances  $\sigma_1^2, \sigma_2^2$  and if

$$W = \frac{X}{\sigma_1}, Z = \frac{1}{\sqrt{(1-\rho^2)}} \left\{ \frac{Y}{\sigma_2} - \frac{\rho X}{\sigma_1} \right\}$$

Show that

$$\frac{\partial(w, z)}{\partial(x, y)} = \frac{1}{\sigma_1 \sigma_2 \sqrt{(1-\rho^2)}}$$

and  $W^2 + Z^2 = \frac{1}{(1 - \rho^2)} \left[ \frac{X^2}{\sigma_1^2} - \frac{2\rho XY}{\sigma_1 \sigma_2} + \frac{Y^2}{\sigma_2^2} \right]$

Deduce that the joint probability differential of  $W$  and  $Z$  is

$$dP = \frac{1}{2\pi} \cdot \exp \left[ -\frac{1}{2} (w^2 + z^2) \right] dw dz$$

and hence that  $W, Z$  are independent normal variates with zero means and unit S.D.'s

[Meerut Univ. M.Sc., 1993]

Hence or otherwise obtain the m.g.f. of the bivariate normal distribution.

22. From a standard bivariate normal population, a random sample of  $n$  observations  $(X_i, Y_i)$ , ( $i = 1, 2, \dots, n$ ) is drawn. Show that the distribution of

$$Z_1 = \frac{1}{n} \sum_{i=1}^n X_i^2 \text{ and } Z_2 = \frac{1}{n} \sum_{i=1}^n Y_i^2$$

has the moment generating function :

$$\text{Constant} \left[ \left( 1 - \frac{2t_1}{n} \right) \left( 1 - \frac{2t_2}{n} \right) - \frac{4\rho^2 t_1 t_2}{n^2} \right]^{-n/2}$$

$$\text{Hint. } M_{Z_1, Z_2}(t_1, t_2) = \left[ E \exp \left( \frac{t_1 x^2}{n} + \frac{t_2 y^2}{n} \right) \right]^n$$

$$= \left\{ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left[ x^2 \left( \frac{t_1}{n} - \frac{1}{2(1 - \rho^2)} \right) + \left( \frac{\rho}{1 - \rho^2} \right) xy \right. \right. \\ \left. \left. + y^2 \left( \frac{t_2}{n} - \frac{1}{2(1 - \rho^2)} \right) \right] dx dy \right\}^n$$

Now use the result

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp[-(ax^2 + 2hxy + by^2)] dx dy = \frac{\pi}{\sqrt{ab - h^2}}$$

and simplify.

**10-11. Multiple and Partial Correlation.** When the values of one variable are associated with or influenced by other variable, e.g., the age of husband and wife, the height of father and son, the supply and demand of a commodity and so on, Karl Pearson's coefficient of correlation can be used as a measure of linear relationship between them. But sometimes there is interrelation between many variables and the value of one variable may be influenced by many others, e.g., the yield of crop per acre say  $(X_1)$  depends upon quality of seed  $(X_2)$ , fertility of soil  $(X_3)$ , fertilizer used  $(X_4)$ , irrigation facilities  $(X_5)$ , weather conditions  $(X_6)$  and so on. Whenever we are interested in studying the joint effect of a group of variables upon a variable not included in that group, our study is that of *multiple correlation and multiple regression*.

Suppose in a trivariate or multi-variate distribution we are interested in the relationship between two variables only. There are two alternatives, viz., (i) we

consider only those two members of the observed data in which the other members have specified values or (ii) we may eliminate mathematically the effect of other variates on two variates. The first method has the disadvantage that it limits the size of the data and also it will be applicable to only the data in which the other variates have assigned values. In the second method it may not be possible to eliminate the entire influence of the variates but the linear effect can be easily eliminated. The correlation and regression between only two variates eliminating the linear effect of other variates in them is called the *partial correlation and partial regression*.

**10.11.1. Yule's Notation.** Let us consider a distribution involving three random variables  $X_1, X_2$  and  $X_3$ . Then the equation of the plane of regression of  $X_1$  on  $X_2$  and  $X_3$  is

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \quad \dots(10.28)$$

Without loss of generality we can assume that the variables  $X_1, X_2$  and  $X_3$  have been measured from their respective means, so that

$$E(X_1) = E(X_2) = E(X_3) = 0$$

Hence on taking expectation of both sides in (10.28), we get  $a = 0$ .

Thus the plane of regression of  $X_1$  on  $X_2$  and  $X_3$  becomes

$$X_1 = b_{12.3}X_2 + b_{13.2}X_3 \quad \dots(10.28a)$$

The coefficients  $b_{12.3}$  and  $b_{13.2}$  are known as the *partial regression coefficients* of  $X_1$  on  $X_2$  and of  $X_1$  on  $X_3$  respectively.

$$e_{1.23} = b_{12.3}X_2 + b_{13.2}X_3$$

is called the estimate of  $X_1$  as given by the plane of regression (10.28a) and the quantity

$$X_{1.23} = X_1 - b_{12.3}X_2 - b_{13.2}X_3,$$

is called the *error of estimate or residual*.

In the general case of  $n$  variables  $X_1, X_2, \dots, X_n$ , the equation of the plane of regression of  $X_1$  on  $X_2, X_3, \dots, X_n$  becomes

$$X_1 = b_{12.34\dots n}X_2 + b_{13.24\dots n}X_3 + \dots + b_{1n.23\dots(n-1)}X_n$$

The error of estimate or residual is given by

$$X_{1.23\dots n} = X_1 - (b_{12.34\dots n}X_2 + b_{13.24\dots n}X_3 + \dots + b_{1n.23\dots(n-1)}X_n)$$

The notations used here are due to Yule. The subscripts before the dot(.) are known as *primary subscripts* and those after the dot are called *secondary subscripts*. The order of a regression coefficient is determined by the number of secondary subscripts, e.g.,

$$b_{12.3}, b_{12.34}, \dots, b_{12.34\dots n}$$

are the regression coefficients of order 1, 2, ...,  $(n - 2)$  respectively. Thus in general, a regression coefficient with  $p$ -secondary subscripts will be called a regression coefficient of order ' $p$ '. It may be pointed out that the order in which the secondary subscripts are written is immaterial but the order of the primary subscripts is important, e.g., in  $b_{12.34\dots n}$ ,  $X_2$  is independent while  $X_1$  is dependent variable but in  $b_{21.34\dots n}$ ,  $X_1$  is independent while  $X_2$  is dependent

variable. Thus of the two primary subscripts, former refers to dependent variable and the latter to independent variable.

The order of a residual is also determined by the number of secondary subscripts in it, e.g.,  $X_{1 \cdot 23}, X_{1 \cdot 234}, \dots, X_{1 \cdot 23 \dots n}$  are the residuals of order 2, 3, ...,  $(n - 1)$  respectively.

**Remark.** In the following sequences we shall assume that the variables under consideration have been measured from their respective means.

**10-12. Plane of Regression.** The equation of the plane of regression of  $X_1$  on  $X_2$  and  $X_3$  is

$$X_1 = b_{12 \cdot 3} X_2 + b_{13 \cdot 2} X_3 \quad \dots(10-29)$$

The constants  $b$ 's in (10-29) are determined by the principle of least squares, i.e., by minimising the sum of the squares of the residuals, viz.,

$$S = \sum X_{1 \cdot 23}^2 = \sum (X_1 - b_{12 \cdot 3} X_2 - b_{13 \cdot 2} X_3)^2,$$

the summation being extended to the given values ( $N$  in number) of the variables.

The normal equations for estimating  $b_{12 \cdot 3}$  and  $b_{13 \cdot 2}$  are

$$\left. \begin{aligned} \frac{\partial S}{\partial b_{12 \cdot 3}} &= 0 = -2 \sum X_2 (X_1 - b_{12 \cdot 3} X_2 - b_{13 \cdot 2} X_3) \\ \frac{\partial S}{\partial b_{13 \cdot 2}} &= 0 = -2 \sum X_3 (X_1 - b_{12 \cdot 3} X_2 - b_{13 \cdot 2} X_3) \end{aligned} \right\} \quad \dots(10-30)$$

$$\text{i.e.,} \quad \sum X_2 X_{1 \cdot 23} = 0 \quad \text{and} \quad \sum X_3 X_{1 \cdot 23} = 0 \quad \dots(10-30a)$$

$$\Rightarrow \left. \begin{aligned} \sum X_1 X_2 - b_{12 \cdot 3} \sum X_2^2 - b_{13 \cdot 2} \sum X_2 X_3 &= 0 \\ \sum X_1 X_3 - b_{12 \cdot 3} \sum X_2 X_3 - b_{13 \cdot 2} \sum X_3^2 &= 0 \end{aligned} \right\} \quad \dots(10-30b)$$

Since  $X_i$ 's are measured from their respective means, we have

$$\left. \begin{aligned} \sigma_i^2 &= \frac{1}{N} \sum X_i^2, \quad \text{Cov}(X_i, X_j) = \frac{1}{N} \sum X_i X_j \\ \text{and} \quad r_{ij} &\doteq \frac{\text{Cov}(X_i, X_j)}{\sigma_i \sigma_j} = \frac{\sum X_i X_j}{N \sigma_i \sigma_j} \end{aligned} \right\} \quad \dots(10-30c)$$

Hence from (10-30b), we get

$$\left. \begin{aligned} r_{12} \sigma_1 \sigma_2 - b_{12 \cdot 3} \sigma_2^2 - b_{13 \cdot 2} r_{23} \sigma_2 \sigma_3 &= 0 \\ r_{13} \sigma_1 \sigma_3 - b_{12 \cdot 3} r_{23} \sigma_2 \sigma_3 - b_{13 \cdot 2} \sigma_3^2 &= 0 \end{aligned} \right\} \quad \dots(10-30d)$$

Solving the equations (10-30d) for  $b_{12 \cdot 3}$  and  $b_{13 \cdot 2}$ , we get

$$b_{12 \cdot 3} = \frac{\begin{vmatrix} r_{12} \sigma_1 & r_{23} \sigma_3 \\ r_{13} \sigma_1 & \sigma_3 \end{vmatrix}}{\begin{vmatrix} \sigma_2 & r_{23} \sigma_3 \\ r_{23} \sigma_2 & \sigma_3 \end{vmatrix}} = \frac{\sigma_1}{\sigma_2} \cdot \frac{\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} \quad \dots(10-31)$$

Similarly, we will get

$$b_{13 \cdot 2} = \frac{\sigma_1}{\sigma_3} \cdot \frac{\begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} \quad \dots(10-31a)$$

If we write

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} \quad \dots(10-32)$$

and  $\omega_{ij}$  is the cofactor of the element in the  $i$ th row and  $j$ th column of  $\omega$ , we have from (10-31) and (10-31a)

$$b_{12 \cdot 3} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \quad \text{and} \quad b_{13 \cdot 2} = -\frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}} \quad \dots(10-33)$$

Substituting these values in (10-29), we get the required equation of the plane of regression of  $X_1$  on  $X_2$  and  $X_3$  as

$$\begin{aligned} X_1 &= -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \cdot X_2 - \frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}} \cdot X_3 \\ \Rightarrow \quad \frac{X_1}{\sigma_1} \cdot \omega_{11} + \frac{X_2}{\sigma_2} \cdot \omega_{12} + \frac{X_3}{\sigma_3} \cdot \omega_{13} &= 0 \end{aligned} \quad \dots(10-34)$$

**Aliter.** Eliminating the coefficient  $b_{12 \cdot 3}$  and  $b_{13 \cdot 2}$  in (10-29) and (10-30a), the required equation of the plane of regression of  $X_1$  on  $X_2$  and  $X_3$  becomes

$$\begin{vmatrix} X_1 & X_2 & X_3 \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & r_{23}\sigma_2\sigma_3 \\ r_{13}\sigma_1\sigma_3 & r_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{vmatrix} = 0$$

Dividing  $C_1$ ,  $C_2$  and  $C_3$  by  $\sigma_1$ ,  $\sigma_2$  and  $\sigma_3$  respectively and also  $R_2$  and  $R_3$  by  $\sigma_2$  and  $\sigma_3$  respectively, we get

$$\begin{aligned} \begin{vmatrix} \frac{X_1}{\sigma_1} & \frac{X_2}{\sigma_2} & \frac{X_3}{\sigma_3} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} &= 0 \\ \Rightarrow \quad \frac{X_1}{\sigma_1} \omega_{11} + \frac{X_2}{\sigma_2} \omega_{12} + \frac{X_3}{\sigma_3} \omega_{13} &= 0 \end{aligned}$$

where  $\omega_{ij}$  is defined in (10-32).

**10-12-1. Generalisation.** In general, the equation of the plane of regression of  $X_1$  on  $X_2, X_3, \dots, X_n$  is

$$X_1 = b_{12 \cdot 34 \dots n} X_2 + b_{13 \cdot 24 \dots n} X_3 + \dots + b_{1n \cdot 23 \dots (n-1)} X_n \quad \dots(10-35)$$

The sum of the squares of residuals is given by

$$S = \sum X_{1 \cdot 23 \dots n}^2$$

$$= \sum (X_1 - b_{12\cdot 34\ldots n} X_2 - b_{13\cdot 24\ldots n} X_3 - \dots - b_{1n\cdot 23\ldots (n-1)} X_n)^2$$

Using the principle of least squares, the normal equations for estimating the  $(n-1)$ ,  $b$ 's are

$$\frac{\partial S}{\partial b_{12\cdot 34\ldots n}} = 0 = -2 \sum X_2 (X_1 - b_{12\cdot 34\ldots n} X_2 - b_{13\cdot 24\ldots n} X_3 - \dots - b_{1n\cdot 23\ldots (n-1)} X_n)$$

$$\frac{\partial S}{\partial b_{13\cdot 24\ldots n}} = 0 = -2 \sum X_3 (X_1 - b_{12\cdot 34\ldots n} X_2 - b_{13\cdot 24\ldots n} X_3 - \dots - b_{1n\cdot 23\ldots (n-1)} X_n)$$

$$\frac{\partial S}{\partial b_{1n\cdot 23\ldots (n-1)}} = 0 = -2 \sum X_n (X_1 - b_{12\cdot 34\ldots n} X_2 - b_{13\cdot 24\ldots n} X_3 - \dots - b_{1n\cdot 23\ldots (n-1)} X_n) \quad \boxed{\dots(10.36)}$$

$$\text{i.e., } \sum X_i X_{1\cdot 23\ldots n} = 0, \quad (i = 2, 3, \dots, n) \quad \boxed{\dots(10.36a)}$$

which on simplification after using (10.30c), give

$$r_{12}\sigma_1\sigma_2 = b_{12\cdot 34\ldots n}\sigma_2^2 + b_{13\cdot 24\ldots n}r_{23}\sigma_2\sigma_3 + \dots + b_{1n\cdot 23\ldots (n-1)}r_{2n}\sigma_2\sigma_n$$

$$r_{13}\sigma_1\sigma_3 = b_{12\cdot 34\ldots n}r_{23}\sigma_2\sigma_3 + b_{13\cdot 24\ldots n}\sigma_3^2 + \dots + b_{1n\cdot 23\ldots (n-1)}r_{3n}\sigma_3\sigma_n$$

$$r_{1n}\sigma_1\sigma_n = b_{12\cdot 34\ldots n}r_{2n}\sigma_2\sigma_n + b_{13\cdot 24\ldots n}r_{3n}\sigma_3\sigma_n + \dots + b_{1n\cdot 23\ldots (n-1)}\sigma_n^2 \quad \boxed{\dots(10.36b)}$$

Hence the eliminant of  $b$ 's between (10.35) and (10.36b) is

$$\left| \begin{array}{ccccc} X_1 & X_2 & X_3 & \dots & X_n \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & r_{23}\sigma_2\sigma_3 & \dots & r_{2n}\sigma_2\sigma_n \\ r_{13}\sigma_1\sigma_3 & r_{23}\sigma_2\sigma_3 & \sigma_3^2 & \dots & r_{3n}\sigma_3\sigma_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1n}\sigma_1\sigma_n & r_{2n}\sigma_2\sigma_n & r_{3n}\sigma_3\sigma_n & \dots & \sigma_n^2 \end{array} \right| = 0$$

Dividing  $C_1, C_2, \dots, C_n$  by  $\sigma_1, \sigma_2, \dots, \sigma_n$  respectively and  $R_2, R_3, \dots, R_n$  by  $\sigma_2, \sigma_3, \dots, \sigma_n$  respectively, we get

$$\left| \begin{array}{ccccc} \frac{X_1}{\sigma_1} & \frac{X_2}{\sigma_2} & \frac{X_3}{\sigma_3} & \dots & \frac{X_n}{\sigma_n} \\ r_{12} & 1 & r_{32} & \dots & r_{2n} \\ r_{13} & r_{23} & 1 & \dots & r_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & r_{3n} & \dots & 1 \end{array} \right| = 0 \quad \boxed{\dots(10.37)}$$

If we write

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ r_{21} & 1 & r_{23} & \dots & r_{2n} \\ r_{31} & r_{32} & 1 & \dots & r_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \dots & 1 \end{vmatrix} \quad \dots(10.38)$$

and  $\omega_{ij}$  is the cofactor of the element in the  $i$ th row and  $j$ th column of  $\omega$ , we get from (10.37)

$$\frac{X_1}{\sigma_1} \cdot \omega_{11} + \frac{X_2}{\sigma_2} \omega_{12} + \frac{X_3}{\sigma_3} \omega_{13} + \dots + \frac{X_n}{\sigma_n} \omega_{1n} = 0 \quad \dots(10.39)$$

as the required equation of the plane of regression of  $X_1$  on  $X_2, X_3, \dots, X_n$ .

Equation (10.39) can be re-written as

$$X_1 = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} X_2 - \frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}} X_3 - \dots - \frac{\sigma_1}{\sigma_n} \cdot \frac{\omega_{1n}}{\omega_{11}} X_n \quad \dots(10.39a)$$

Comparing (10.39a) with (10.35), we get

$$\left. \begin{aligned} b_{12.34\dots n} &= -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}}, \\ b_{13.24\dots n} &= -\frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}}, \\ &\vdots &&\vdots \\ b_{1n.23\dots(n-1)} &= -\frac{\sigma_1}{\sigma_n} \cdot \frac{\omega_{1n}}{\omega_{11}} \end{aligned} \right\} \quad \dots(10.40)$$

**Remarks 1.** From the symmetry of the result obtained in (10.40), the equation of the plane of regression of  $X_i$ , (say), on the remaining variables  $X_j$  ( $j \neq i = 1, 2, \dots, n$ ), is given by

$$\frac{X_1}{\sigma_1} \omega_{i1} + \frac{X_2}{\sigma_2} \omega_{i2} + \dots + \frac{X_i}{\sigma_i} \omega_{ii} + \dots + \frac{X_n}{\sigma_n} \omega_{in} = 0 ; i = 1, 2, \dots, n \quad \dots(10.41)$$

**2. We have**

$$b_{12.34\dots n} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}}$$

$$\text{and} \quad b_{21.34\dots n} = -\frac{\sigma_2}{\sigma_1} \cdot \frac{\omega_{21}}{\omega_{22}}$$

Since each of  $\sigma_1, \sigma_2, \omega_{11}$  and  $\omega_{22}$  is non-negative and  $\omega_{12} = \omega_{21}$ , [c.f. Remarks 3 and 4 to §10.14, page 10.113], the sign of each regression coefficient  $b_{12.34\dots n}$  and  $b_{21.34\dots n}$  depends on  $\omega_{12}$ .

### 10-13. Properties of residuals

**Property 1.** *The sum of the product of any residual of order zero with any other residual of higher order is zero, provided the subscript of the former occurs among the secondary subscripts of the latter.*

The normal equations for estimating  $b$ 's in trivariate and  $n$ -variate distributions, as obtained in equations (10-30a) and (10-36a), are

$$\sum X_2 X_{1 \cdot 23} = 0, \sum X_3 X_{1 \cdot 23} = 0$$

and

$$\sum X_i X_{1 \cdot 23 \dots n} = 0; i = 2, 3, \dots, n$$

respectively. Here  $X_i$ , ( $i = 1, 2, 3, \dots, n$ ) can be regarded as a residual of order zero. Hence the result.

**Property 2.** *The sum of the product of any two residuals in which all the secondary subscripts of the first occur among the secondary subscripts of the second is unaltered if we omit any or all of the secondary subscripts of the first. Conversely, the product sum of any residual of order ' $p$ ' with a residual of order  $p + q$ , the ' $p$ ' subscripts being the same in each case is unaltered by adding to the secondary subscripts of the former any or all the ' $q$ ' additional subscripts of the latter.*

Let us consider

$$\begin{aligned}\sum X_{1 \cdot 2} X_{1 \cdot 23} &= \sum (X_1 - b_{12} X_2) X_{1 \cdot 23} = \sum X_1 X_{1 \cdot 23} - b_{12} \sum X_2 X_{1 \cdot 23} \\ &= \sum X_1 X_{1 \cdot 23} \quad (\text{c.f. Property 1})\end{aligned}$$

$$\begin{aligned}\text{Also } \sum X_{1 \cdot 23}^2 &= \sum X_{1 \cdot 23} X_{1 \cdot 23} = \sum (X_1 - b_{12 \cdot 3} X_2 - b_{13 \cdot 2} X_3) X_{1 \cdot 23} \\ &= \sum X_1 X_{1 \cdot 23} - b_{12 \cdot 3} \sum X_2 X_{1 \cdot 23} - b_{13 \cdot 2} \sum X_3 X_{1 \cdot 23} \\ &= \sum X_1 X_{1 \cdot 23} \quad (\text{c.f. Property 1})\end{aligned}$$

$$\therefore \sum X_{1 \cdot 23}^2 = \sum X_{1 \cdot 2} X_{1 \cdot 23} = \sum X_1 X_{1 \cdot 23}$$

Again  $\sum X_{1 \cdot 34 \dots n} X_{2 \cdot 34 \dots n}$

$$\begin{aligned}&= \sum [(X_1 - b_{13 \cdot 4 \dots n} X_3 - b_{14 \cdot 35 \dots n} X_4 - \dots - b_{1n \cdot 34 \dots (n-1)} X_n) X_{2 \cdot 34 \dots n}] \\ &= \sum X_1 X_{2 \cdot 34 \dots n} \quad (\text{c.f. Property 1})\end{aligned}$$

Hence the property ?

**Property 3.** *The sum of the product of two residuals is zero if all the subscripts (primary as well as secondary) of the one occur among the secondary subscripts of the other, e.g.,*

$$\sum X_{1 \cdot 2} X_{3 \cdot 12} = \sum (X_1 - b_{12} X_2) X_{3 \cdot 12} = \sum X_1 X_{3 \cdot 12} - b_{12} \sum X_2 X_{3 \cdot 12} = 0 \quad (\text{c.f. Property 1})$$

$$\sum X_{2 \cdot 34 \dots n} X_{1 \cdot 23 \dots n}$$

$$\begin{aligned}&= \sum [(X_2 - b_{23 \cdot 4 \dots n} X_3 - b_{24 \cdot 35 \dots n} X_4 - \dots - b_{2n \cdot 34 \dots (n-1)} X_n) X_{1 \cdot 23 \dots n}] \\ &= \sum X_2 X_{1 \cdot 23 \dots n} - b_{23 \cdot 4 \dots n} \sum X_3 X_{1 \cdot 23 \dots n} - b_{24 \cdot 35 \dots n} \sum X_4 X_{1 \cdot 23 \dots n} \\ &\quad \dots - b_{2n \cdot 34 \dots (n-1)} \sum X_n X_{1 \cdot 23 \dots n} \\ &= 0 \quad (\text{c.f. Property 1})\end{aligned}$$

Hence the property 3.

**10.13.1. Variance of the Residual.** Let us consider the plane of regression of  $X_1$  on  $X_2, X_3, \dots, X_n$  viz.,

$$X_1 = b_{123\dots n} X_2 + b_{132\dots n} X_3 + \dots + b_{1n23\dots(n-1)} X_n$$

Since all the  $X_i$ 's are measured from their respective means, we have

$$E(X_i) = 0; i = 1, 2, \dots, n \Rightarrow E(X_{123\dots n}) = 0$$

Hence the variance of the residual is given by

$$\begin{aligned}\sigma^2_{123\dots n} &= \frac{1}{N} \sum [X_{123\dots n} - E(X_{123\dots n})]^2 = \frac{1}{N} \sum X_{123\dots n}^2 \\ &= \frac{1}{N} \sum X_{123\dots n} X_{123\dots n} = \frac{1}{N} \sum X_1 X_{123\dots n},\end{aligned}$$

(c.f. Property 2 § 10.13)

$$\begin{aligned}&= \frac{1}{N} \sum X_1 (X_1 - b_{123\dots n} X_2 - b_{132\dots n} X_3 - \dots - b_{1n23\dots(n-1)} X_n) \\ &= \sigma_1^2 - b_{123\dots n} r_{12} \sigma_1 \sigma_2 - b_{132\dots n} r_{13} \sigma_1 \sigma_3 - \dots - b_{1n23\dots(n-1)} r_{1n} \sigma_1 \sigma_n \\ \Rightarrow \quad \sigma_1^2 - \sigma^2_{123\dots n} &= b_{123\dots n} r_{12} \sigma_1 \sigma_2 - b_{132\dots n} r_{13} \sigma_1 \sigma_3 - \dots \\ &\quad - b_{1n23\dots(n-1)} r_{1n} \sigma_1 \sigma_n \dots \quad (10.42)\end{aligned}$$

Eliminating the  $b$ 's in equations (10.42) and (10.36b), we get

$$\left| \begin{array}{cccc} \sigma_1^2 - \sigma^2_{123\dots n} & r_{12} \sigma_1 \sigma_2 & \dots & r_{1n} \sigma_1 \sigma_n \\ r_{12} \sigma_1 \sigma_2 & \sigma_2^2 & \dots & r_{2n} \sigma_2 \sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} \sigma_1 \sigma_n & r_{2n} \sigma_2 \sigma_n & \dots & \sigma_n^2 \end{array} \right| = 0$$

Dividing  $R_1, R_2, \dots, R_n$ , by  $\sigma_1, \sigma_2, \dots, \sigma_n$  respectively and also  $C_1, C_2, \dots, C_n$  by  $\sigma_1, \sigma_2, \dots, \sigma_n$  respectively, we get

$$\begin{aligned}&\left| \begin{array}{cccc} 1 - \frac{\sigma^2_{123\dots n}}{\sigma_1^2} & r_{12} & \dots & r_{1n} \\ r_{12} & 1 & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \dots & 1 \end{array} \right| = 0 \\ \Rightarrow \quad &\left| \begin{array}{cccc} 1 - \frac{\sigma^2_{123\dots n}}{\sigma_1^2} & r_{12} & \dots & r_{1n} \\ r_{12} + 0 & 1 & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} + 0 & r_{2n} & \dots & 1 \end{array} \right| = 0\end{aligned}$$

$$\left| \begin{array}{cccc} 1 & r_{12} & \dots & r_{1n} \\ r_{12} & 1 & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \dots & 1 \end{array} \right| - \left| \begin{array}{cccc} \frac{\sigma^2_{1.23\dots n}}{\sigma_1^2} & r_{12} & \dots & r_{1n} \\ 0 & 1 & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & r_{2n} & \dots & 1 \end{array} \right| = 0$$

$$\Rightarrow \omega - \frac{\sigma^2_{1.23\dots n}}{\sigma_1^2} \omega_{11} = 0$$

$$\therefore \sigma^2_{1.23\dots n} = \sigma_1^2 \frac{\omega}{\omega_{11}} \quad \dots(10.43)$$

**Remark.** In a tri-variate distribution,

$$\sigma_{1.23}^2 = \sigma_1^2 \frac{\omega}{\omega_{11}} \quad \dots(10.43a)$$

where  $\omega$  and  $\omega_{11}$  are defined in (10.32).

**10.14. Coefficient of Multiple Correlation.** In a tri-variate distribution in which each of the variables  $X_1$ ,  $X_2$ , and  $X_3$  has  $N$  observations, the multiple correlation coefficient of  $X_1$  on  $X_2$  and  $X_3$ , usually denoted by  $R_{1.23}$ , is the simple correlation coefficient between  $X_1$  and the joint effect of  $X_2$  and  $X_3$  on  $X_1$ . In other words  $R_{1.23}$  is the correlation coefficient between  $X_1$  and its estimated value as given by the plane of regression of  $X_1$  on  $X_2$  and  $X_3$  viz.,

$$e_{1.23} = b_{12.3} X_2 + b_{13.2} X_3$$

$$\text{We have } X_{1.23} = X_1 - b_{12.3} X_2 - b_{13.2} X_3 = X_1 - e_{1.23}$$

$$\Rightarrow e_{1.23} = X_1 - X_{1.23}$$

Since  $X_i$ 's are measured from their respective means, we have

$$E(X_{1.23}) = 0 \text{ and } E(e_{1.23}) = 0 \quad (\because E(X_i) = 0; i = 1, 2, 3)$$

By def.,

$$R_{1.23} = \frac{\text{Cov}(X_1, e_{1.23})}{\sqrt{V(X_1) V(e_{1.23})}} \quad \dots(10.44)$$

$$\begin{aligned} \text{Cov}(X_1, e_{1.23}) &= E[(X_1 - E(X_1))(e_{1.23} - E(e_{1.23}))] = E(X_1 e_{1.23}) \\ &= \frac{1}{N} \sum X_1 e_{1.23} = \frac{1}{N} \sum X_1 (X_1 - X_{1.23}) \\ &= \frac{1}{N} \sum X_1^2 - \frac{1}{N} \sum X_1 X_{1.23} = \frac{1}{N} \sum X_1^2 - \frac{1}{N} \sum X_{1.23}^2 \\ &= \sigma_1^2 - \sigma_{1.23}^2 \quad (\text{c.f. Property 2, § 10.13}) \end{aligned}$$

$$\begin{aligned} \text{Also } V(e_{1.23}) &= E(e_{1.23}^2) = \frac{1}{N} \sum e_{1.23}^2 = \frac{1}{N} \sum (X_1 - X_{1.23})^2 \\ &= \frac{1}{N} \sum (X_1^2 + X_{1.23}^2 - 2 X_1 X_{1.23}) \\ &= \frac{1}{N} \sum X_1^2 + \frac{1}{N} \sum X_{1.23}^2 - \frac{2}{N} \sum X_1 X_{1.23} \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{N} \sum X_1^2 + \frac{1}{N} \sum X_{1-23}^2 - \frac{2}{N} \sum X_{1-23} \\
 &= \sigma_1^2 - \sigma_{1-23}^2 \quad (\text{cf. Property 2, § 10.13}) \\
 \therefore R_{1-23} &= \frac{\sigma_1^2 - \sigma_{1-23}^2}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma_{1-23}^2)}} \\
 \Rightarrow R_{1-23}^2 &= \frac{\sigma_1^2 - \sigma_{1-23}^2}{\sigma_1^2} = 1 - \frac{\sigma_{1-23}^2}{\sigma_1^2} \\
 \Rightarrow 1 - R_{1-23}^2 &= \frac{\sigma_{1-23}^2}{\sigma_1^2}
 \end{aligned}$$

Using (10.43a), we get

$$1 - R_{1-23}^2 = \frac{\omega}{\omega_{11}} \quad \dots(10.45)$$

where

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} \quad (\text{On simplification}).$$

$$\text{and } \omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2$$

Hence from (10.45), we get

$$R_{1-23}^2 = 1 - \frac{\omega}{\omega_{11}} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} \quad \dots(10.45a)$$

This formula expresses the multiple correlation coefficient in terms of the total correlation coefficients between the pairs of variables.

**Generalisation.** In case of  $n$ -variate distribution, the multiple correlation coefficient of  $X_1$  on  $X_2, X_3, \dots, X_n$ , usually denoted by  $R_{1-23\dots n}$ , is the correlation coefficient between  $X_1$  and

$$\begin{aligned}
 e_{1-23\dots n} &= X_1 - X_{1-23\dots n} \\
 \therefore R_{1-23\dots n} &= \frac{\text{Cov}(X_1, e_{1-23\dots n})}{\sqrt{V(X_1) V(e_{1-23\dots n})}} \\
 \text{Cov}(X_1, e_{1-23\dots n}) &= \frac{1}{N} \sum X_1 e_{1-23\dots n} = \frac{1}{N} \sum X_1 (X_1 - X_{1-23\dots n}) \\
 &= \frac{1}{N} \sum X_1^2 - \frac{1}{N} \sum X_1 X_{1-23\dots n} \\
 &= \frac{1}{N} \sum X_1^2 - \frac{1}{N} \sum X_{1-23\dots n}^2 = \sigma_1^2 - \sigma_{1-23\dots n}^2 \quad \dots(*) \\
 V(e_{1-23\dots n}) &= \frac{1}{N} \sum e_{1-23\dots n}^2 = \frac{1}{N} \sum (X_1 - X_{1-23\dots n})^2
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{N} \sum (X_1^2 + X_{1.23...n}^2 - 2X_1 X_{1.23...n}) \\
 &= \frac{1}{N} \sum X_1^2 + \frac{1}{N} \sum X_{1.23...n}^2 - 2 \frac{1}{N} \sum X_1 X_{1.23...n} \\
 &= \frac{1}{N} \sum X_1^2 + \frac{1}{N} \sum X_{1.23...n}^2 - \frac{2}{N} \sum X_{1.23...n}^2 \\
 &= \sigma_1^2 - \sigma_{1.23...n}^2 \\
 \therefore R_{1.23...n} &= \frac{\sigma_1^2 - \sigma_{1.23...n}^2}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma_{1.23...n}^2)}} = \left( \frac{\sigma_1^2 - \sigma_{1.23...n}^2}{\sigma_1^2} \right)^{1/2} \\
 R_{1.23...n}^2 &= 1 - \frac{\sigma_{1.23...n}^2}{\sigma_1^2} = 1 - \frac{\omega}{\omega_{11}} \quad \dots(10.45c)
 \end{aligned}$$

where  $\omega$  and  $\omega_{11}$  are defined in (10.38).

**Remarks 1.** It may be pointed out here that multiple correlation coefficient can never be negative, because from (\*) and (\*\*), we get

$$\text{Cov}(X_1, e_{1.23...n}) = \sigma_1^2 - \sigma_{1.23...n}^2 = \text{Var}(e_{1.23...n}) \geq 0$$

Since the sign of  $R_{1.23...n}$  depends upon the covariance term  $\text{Cov}(X_1, e_{1.23...n})$ , we conclude that  $R_{1.23...n} \geq 0$ .

2. Since  $R_{1.23...n}^2 \geq 0$ , we have :

$$1 - \frac{\omega}{\omega_{11}} \geq 0 \Rightarrow \omega \leq \omega_{11} \quad \dots(10.45d)$$

$$3. \text{ Also, } R_{1.23...n}^2 \leq 1 \Rightarrow 1 - \frac{\omega}{\omega_{11}} \leq 1$$

$$\Rightarrow 0 \leq \frac{\omega}{\omega_{11}} \Rightarrow \frac{\omega}{\omega_{11}} \geq 0 \Rightarrow \omega \geq 0 \quad \dots(10.45e)$$

From the above results, we get

$$\omega_{11} \geq \omega \geq 0 \quad \dots(10.45f)$$

In general, we have

$$\omega_{ii} \geq 0; i = 1, 2, \dots, n$$

4. Since  $\omega$  is symmetric in  $r_{ij}$ 's, we have

$$\omega_{ij} = \omega_{ji}; i \neq j = 1, 2, \dots, n \quad \dots(10.45g)$$

#### 10.14.1. Properties of Multiple Correlation Coefficient

1. Multiple correlation co-efficient measures the closeness of the association between the observed values and the expected values of a variable obtained from the multiple linear regression of that variable on other variables.

2. Multiple correlation coefficient between observed values and expected values, when the expected values are calculated from a linear relation of the variables determined by the method of least squares, is always greater than that where expected values are calculated from any other linear combination of the variables.

3. Since  $R_{1.23}$  is the simple correlation between  $X_1$  and  $e_{1.23}$ , it must lie between -1 and +1. But as seen in Remark 1 above,  $R_{1.23}$  is a non-negative quantity and we conclude that  $0 \leq R_{1.23} \leq 1$ .

4. If  $R_{1.23} = 1$ , then association is perfect and all the regression residuals are zero, and as such  $\sigma^2_{1.23} = 0$ . In this case, since  $X_1 = e_{1.23}$ , the predicted value of  $X_1$ , the multiple linear regression equation of  $X_1$  on  $X_2$  and  $X_3$  may be said to be a perfect prediction formula.

5. If  $R_{1.23} = 0$ , then all total and partial correlations involving  $X_1$  are zero. [See Example 10.37]. So  $X_1$  is completely uncorrelated with all the other variables in this case and the multiple regression equation fails to throw any light on the value of  $X_1$  when  $X_2$  and  $X_3$  are known.

6.  $R_{1.23}$  is not less than any total correlation coefficient, i.e.,

$$R_{1.23} \geq r_{12}, r_{13}, r_{23}$$

**10.15. Coefficient of Partial Correlation.** Sometimes the correlation between two variables  $X_1$  and  $X_2$  may be partly due to the correlation of a third variable,  $X_3$  with both  $X_1$  and  $X_2$ . In such a situation, one may want to know what the correlation between  $X_1$  and  $X_2$  would be if the effect of  $X_3$  on each of  $X_1$  and  $X_2$  were eliminated. This correlation is called the *partial correlation* and the correlation coefficient between  $X_1$  and  $X_2$  after the linear effect of  $X_3$  on each of them has been eliminated is called the *partial correlation coefficient*.

The residual  $X_{1.3} = X_1 - b_{13}X_3$ , may be regarded as that part of the variable  $X_1$  which remains after the linear effect of  $X_3$  has been eliminated. Similarly, the residual  $X_{2.3}$  may be interpreted as the part of the variable  $X_2$  obtained after eliminating the linear effect of  $X_3$ . Thus the partial correlation coefficient between  $X_1$  and  $X_2$ , usually denoted by  $r_{12.3}$ , is given by

$$r_{12.3} = \frac{\text{Cov}(X_{1.3}, X_{2.3})}{\sqrt{\text{Var}(X_{1.3}) \text{Var}(X_{2.3})}} \quad \dots(10.46)$$

We have

$$\begin{aligned} \text{Cov}(X_{1.3}, X_{2.3}) &= \frac{1}{N} \sum X_{1.3} X_{2.3} = \frac{1}{N} \sum X_1 X_{2.3} \\ &= \frac{1}{N} \sum X_1 (X_2 - b_{23} X_3) = \frac{1}{N} \sum X_1 X_2 - b_{23} \frac{1}{N} \sum X_1 X_3 \\ &= r_{12} \sigma_1 \sigma_2 - r_{23} \frac{\sigma_2}{\sigma_3} \cdot (r_{13} \sigma_1 \sigma_3). \\ &= \sigma_1 \sigma_2 (r_{12} - r_{13} r_{23}) \end{aligned}$$

$$\text{Also } V(X_{1.3}) = \frac{1}{N} \sum X_{1.3}^2 = \frac{1}{N} \sum X_{1.3} X_{1.3}$$

$$= \frac{1}{N} \sum X_1 X_{1.3} = \frac{1}{N} \sum X_1 (X_1 - b_{13} X_3)$$

$$\begin{aligned}
 &= \frac{1}{N} \sum X_1^2 - b_{13} \cdot \frac{1}{N} \sum X_1 X_3 \\
 &= \sigma_1^2 - r_{13} \frac{\sigma_1}{\sigma_3} r_{13} \sigma_1 \sigma_3 \\
 &= \sigma_1^2 (1 - r_{13}^2)
 \end{aligned}$$

Similarly, we shall get

$$V(X_{2,3}) = \sigma_2^2 (1 - r_{23}^2)$$

Hence

$$r_{12,3} = \frac{\sigma_1 \sigma_2 (r_{12} - r_{13} r_{23})}{\sqrt{\sigma_1^2 (1 - r_{13}^2) \sigma_2^2 (1 - r_{23}^2)}} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad \dots(10-46a)$$

Aliter. We have

$$\begin{aligned}
 0 &= \sum X_{2,3} X_{1,3} \\
 &= \sum X_{2,3} (X_1 - b_{12,3} X_2 - b_{13,2} X_3) \\
 &= \sum X_1 X_{2,3} - b_{12,3} \sum X_{2,3} X_2 - b_{13,2} \sum X_{2,3} X_3 \\
 &= \sum X_{1,3} X_{2,3} - b_{12,3} \sum X_{2,3} X_{2,3} \\
 \therefore b_{12,3} &= \frac{\sum X_{1,3} X_{2,3}}{\sum X_{2,3}^2}.
 \end{aligned}$$

From this it follows that  $b_{12,3}$  is coefficient of regression of  $X_{1,3}$  on  $X_{2,3}$ .

Similarly,  $b_{21,3}$  is the coefficient of regression of  $X_{2,3}$  on  $X_{1,3}$ .

Since correlation coefficient is the geometric mean between regression coefficients, we have

$$r^2_{12,3} = b_{12,3} \times b_{21,3}$$

But by def.,

$$\begin{aligned}
 b_{12,3} &= -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \quad \text{and} \quad b_{21,3} = -\frac{\sigma_2}{\sigma_1} \cdot \frac{\omega_{21}}{\omega_{22}} \\
 \therefore r^2_{12,3} &= \left( -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \right) \left( -\frac{\sigma_2}{\sigma_1} \cdot \frac{\omega_{21}}{\omega_{22}} \right) = \frac{\omega_{12}^2}{\omega_{11} \omega_{22}} \\
 &\quad (\because \omega_{12} = \omega_{21}) \\
 \Rightarrow r_{12,3} &= -\frac{\omega_{12}}{\sqrt{\omega_{11} \omega_{22}}},
 \end{aligned}$$

the negative sign being taken since the sign of regression coefficients is the same as that of  $(-\omega_{12})$ .

Substituting the values of  $\omega_{12}$ ,  $\omega_{11}$  and  $\omega_{22}$  from (10-32), we get

$$r_{12,3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

**Remarks 1.** The expressions for  $r_{13,2}$  and  $r_{23,1}$  can be similarly obtained, to give

$$r_{13.2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} \quad \text{and} \quad r_{23.1} = \frac{r_{23} - r_{21} r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}}$$

2. If  $r_{12.3} = 0$ , we have then  $r_{12} = r_{13} r_{23}$ , it means that  $r_{12}$  will not be zero if  $X_3$  is correlated with both  $X_1$  and  $X_2$ . Thus, although  $X_1$  and  $X_2$  may be uncorrelated when effect of  $X_3$  is eliminated, yet  $X_1$  and  $X_2$  may appear to be correlated because they carry the effect of  $X_3$  on them.

3. Partial correlation coefficient helps in deciding whether to include or not an additional independent variable in regression analysis.

4. We know that  $\sigma_1^2(1 - r_{12}^2)$  and  $\sigma_1^2(1 - r_{13}^2)$  are the residual variances if  $X_1$  is estimated from  $X_2$  and  $X_3$  individually, while  $\sigma_1^2(1 - R_{1.23}^2)$  is the residual variance if  $X_1$  is estimated from  $X_2$  and  $X_3$  taken together. So from the above remark and  $R_{1.23}^2 \geq r_{12}^2$  and  $r_{13}^2$ , it follows that inclusion of an additional variable can only reduce the residual variance. Now inclusion of  $X_3$  when  $X_2$  has already been taken for predicting  $X_1$ , is worthwhile only when the resultant reduction in the residual variance is substantial. This will be the case when  $r_{13.2}$  is sufficiently large. Thus in this respect partial correlation coefficient has its significance in regression analysis.

**10-15-1. Generalisation.** In the case of  $n$  variables  $X_1, X_2, \dots, X_n$  the partial correlation coefficient  $r_{12.34\dots n}$  between  $X_1$  and  $X_2$  (after the linear effect of  $X_3, X_4, \dots, X_n$  on them has been eliminated), is given by

$$r_{12.34\dots n}^2 = b_{12.34\dots n} \times b_{21.34\dots n}$$

But, we have

$$\begin{aligned} \text{and} \quad b_{12.34\dots n} &= -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \\ b_{21.34\dots n} &= -\frac{\sigma_2}{\sigma_1} \cdot \frac{\omega_{21}}{\omega_{22}} \end{aligned} \left. \right\} [\text{cf. Equation (10-40)}].$$

$$\therefore r_{12.34\dots n}^2 = \left( -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \right) \left( -\frac{\sigma_2}{\sigma_1} \cdot \frac{\omega_{21}}{\omega_{22}} \right) = \frac{\omega_{12}^2}{\omega_{11}\omega_{22}}$$

$$\Rightarrow r_{12.34\dots n} = -\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}} \quad (10-46b)$$

negative sign being taken since the sign of the regression coefficient is same as that of  $(-\omega_{12})$ .

#### 10-16. Multiple Correlation in Terms of Total and Partial Correlations.

$$1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2) \quad \dots (10-46c)$$

**Proof.** We have

$$\begin{aligned} 1 - R_{1.23}^2 &= 1 - \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} \\ &= \frac{1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} \end{aligned}$$

Also

$$1 - r_{13.2}^2 = 1 - \frac{(r_{13} - r_{12} r_{23})^2}{(1 - r_{12}^2)(1 - r_{23}^2)} = \frac{1 - r_{12}^2 - r_{23}^2 - r_{13}^2 + 2r_{12}r_{13}}{(1 - r_{12}^2)(1 - r_{23}^2)}$$

Hence the result.

**Theorem.** Any standard deviation of order 'p' may be expressed in terms of a standard deviation of order  $(p - 1)$  and a partial correlation coefficient of order  $(p - 1)$ .

**Proof.** Let us consider the sum :

$$\begin{aligned} \sum X_{1.23...n}^2 &= \sum X_{1.23...n} X_{1.23...n} \\ &= \sum [X_{1.23...(n-1)} X_{1.23...n} \\ &\quad (c.f. \text{ Property 2, § 10.13}) \\ &= \sum [X_{1.23...(n-1)} (X_1 - b_{12.34...n} X_2 - \dots - b_{1(n-1).23...n} X_{n-1} \\ &\quad - b_{1n.23...(n-1)} X_n)] \\ &= \sum X_{1.23.(n-1)} X_1 - b_{1n.23...(n-1)} \sum X_{1.23...(n-1)} X_n \\ &\quad (c.f. \text{ Property 2 § 10.13}) \\ &= \sum X_{1.23...(n-1)}^2 - b_{1n.23...(n-1)} \sum X_{1.23...(n-1)} X_{n.23...(n-1)} \end{aligned}$$

Dividing both sides by  $N$  (total number of observations), we get  $\sigma_{1.23...n}^2 = \sigma_{1.23...(n-1)}^2 - b_{1n.23...(n-1)} \text{Cov}(X_{1.23...(n-1)}, X_{n.23...(n-1)})$

The regression coefficient of  $X_{n.23...(n-1)}$  on  $X_{1.23...(n-1)}$  is given by

$$\begin{aligned} b_{n1.23...(n-1)} &= \frac{\text{Cov}(X_{1.23...(n-1)}, X_{n.23...(n-1)})}{\sigma_{1.23...(n-1)}^2}, \\ \therefore \sigma_{1.23...n}^2 &= \sigma_{1.23...(n-1)}^2 [1 - b_{1n.23...(n-1)} b_{n1.23...(n-1)}] \\ &= \sigma_{1.23...(n-1)}^2 [1 - r_{1n.23...(n-1)}^2], \quad \dots(10.47) \end{aligned}$$

a formula which expresses the standard deviation of order  $(n - 1)$  in terms of standard deviation of order  $(n - 2)$  and partial correlation coefficient of order  $(n - 2)$ . If we take  $p = (n - 1)$ , the theorem is established.

**Cor. 1.** From (10.47), we have

$$\sigma_{1.23...(n-1)}^2 = \sigma_{1.23...(n-2)}^2 (1 - r_{1(n-1).23...(n-2)}^2) \quad \dots(10.47a)$$

and so on. Thus the repeated application of (10.47) gives

$$\sigma_{1.23...n}^2 = \sigma_1^2 (1 - r_{12}^2) (1 - r_{13.2}^2) (1 - r_{14.32}^2) \dots (1 - r_{1n.23...(n-1)}^2) \quad \dots(10.47b)$$

Since partial correlation coefficients cannot exceed unity numerically, we get from (10.47), (10.47a), and so on,

$$\begin{array}{lcl}
 \sigma_{1,23\dots n}^2 & \leq & \sigma_{1,23\dots (n-1)}^2 \\
 \sigma_{1,23\dots (n-1)}^2 & \leq & \sigma_{1,23\dots (n-2)}^2 \\
 | & & | \\
 \sigma_{1,23}^2 & \leq & \sigma_{1,2}^2 \\
 \sigma_{1,2}^2 & \leq & \sigma_1^2
 \end{array} \quad \left. \right\} \quad \Rightarrow \quad \sigma_1 \geq \sigma_{1,2} \geq \sigma_{1,23} \geq \dots \geq \sigma_{1,23\dots n} \quad \dots(10-47c)$$

**Cor. 2.** Also, we have

$$\sigma_{1,23\dots n}^2 = \sigma_1^2(1 - R_{1,23\dots n}^2)$$

On using (10-47b), we get

$$1 - R_{1,23\dots n}^2 = (1 - r_{12}^2)(1 - r_{13,2}^2) \dots (1 - r_{1,n-3\dots (n-1)}^2) \quad \dots(10-47d)$$

This is the generalisation of the result obtained in (10-46c).

Since  $|r_{ij,(s)}| \leq 1$ ;  $s = 0, 1, 2, \dots, (n-1)$ ,

where  $r_{ij,(s)}$  is a partial correlation coefficient of order  $s$ . we get from (10-47d)

$$1 - R_{1,23\dots n}^2 \leq 1 - r_{12}^2$$

$$1 - R_{1,23\dots n}^2 \leq 1 - r_{13,2}^2,$$

and so on.

$$i.e., \quad R_{1,23\dots n}^2 \geq r_{12}^2, r_{13,2}^2, \dots, r_{1,n-3\dots (n-1)}^2 \quad \dots(10-47e)$$

Since  $R_{1,23\dots n}$  is symmetric in its secondary subscripts, we have

$$\begin{array}{l}
 R_{1,23\dots n}^2 \geq r_{1i}^2, (i = 2, 3, \dots, n) \\
 R_{1,23\dots n}^2 \geq r_{1i,j} (i \neq j = 2, 3, \dots, n)
 \end{array} \quad \left. \right\} \quad \dots(10-47f)$$

and so on

**10-17. Expression for Regression Coefficients in Terms of Regression Coefficients of Lower Order.** Consider-

$$\begin{aligned}
 \sum X_{1,34\dots n} X_{2,34\dots n} &= \sum X_{1,34\dots (n-1)} X_{2,34\dots n} \\
 &= \sum X_{1,34\dots (n-1)} (X_2 - b_{23,4\dots n} X_3 - \dots - b_{2,n-34\dots (n-1)} X_n) \\
 &= \sum X_{1,34\dots (n-1)} X_2 - b_{2,n-34\dots (n-1)} \sum X_{1,34\dots (n-1)} X_n \\
 &= \sum X_{1,34\dots (n-1)} X_{2,34\dots (n-1)} \\
 &\quad - b_{2,n-34\dots (n-1)} \sum X_{1,34\dots (n-1)} X_{n-34\dots (n-1)}
 \end{aligned}$$

Dividing both sides by  $N$ , the total number of observations, we get

$$\begin{aligned}
 \text{Cov}(X_{1,34\dots n}, X_{2,34\dots n}) &= \text{Cov}(X_{1,34\dots (n-1)}, X_{2,34\dots (n-1)}) \\
 &\quad - b_{2,n-34\dots (n-1)} \text{Cov}(X_{1,34\dots (n-1)}, X_{n-34\dots (n-1)}) \\
 b_{12,34\dots n} \sigma_{2,34\dots n}^2 &= b_{12,34\dots (n-1)} \sigma_{2,34\dots (n-1)}^2 \\
 &\quad - b_{2,n-34\dots (n-1)} b_{1n-34\dots (n-1)} \sigma_{n-34\dots (n-1)}^2
 \end{aligned}$$

On using (10-47), we get

$$\begin{aligned}
 b_{12 \cdot 34 \dots n} \sigma_{2 \cdot 34 \dots (n-1)}^2 & \{1 - r_{2 \cdot 34 \dots (n-1)}^2\} \\
 & = \sigma_{2 \cdot 34 \dots (n-1)}^2 [b_{12 \cdot 34 \dots (n-1)} - b_{2 \cdot 34 \dots (n-1)} b_{1 \cdot 34 \dots (n-1)}] \\
 & \quad \times \left[ \frac{\sigma_{1 \cdot 34 \dots (n-1)}^2}{\sigma_{2 \cdot 34 \dots (n-1)}^2} \right] \dots (*)
 \end{aligned}$$

In the case of two variables, we have

$$b_{ij} \sigma_j^2 = b_{ji} \sigma_i^2 = \text{Cov}(X_i, X_j) \Rightarrow b_{ij} = \frac{\sigma_i^2}{\sigma_j^2} b_{ji}$$

$$\therefore b_{2 \cdot 34 \dots (n-1)} \frac{\sigma_{n \cdot 34 \dots (n-1)}^2}{\sigma_{2 \cdot 34 \dots (n-1)}^2} = b_{n \cdot 34 \dots (n-1)}$$

Hence from (\*), we get

$$\begin{aligned}
 b_{12 \cdot 34 \dots n} \sigma_{2 \cdot 34 \dots (n-1)}^2 & \{1 - r_{2 \cdot 34 \dots (n-1)}^2\} \\
 & = \sigma_{2 \cdot 34 \dots (n-1)}^2 [b_{12 \cdot 34 \dots (n-1)} - b_{1 \cdot 34 \dots (n-1)} b_{n \cdot 34 \dots (n-1)}] \\
 \therefore b_{12 \cdot 34 \dots n} & = \left[ \frac{b_{12 \cdot 34 \dots (n-1)} - b_{1 \cdot 34 \dots (n-1)} b_{n \cdot 34 \dots (n-1)}}{1 - r_{2 \cdot 34 \dots (n-1)}^2} \right] \dots (10.48)
 \end{aligned}$$

$$\Rightarrow b_{12 \cdot 34 \dots n} = \left[ \frac{b_{12 \cdot 34 \dots (n-1)} - b_{1 \cdot 34 \dots (n-1)} b_{n \cdot 34 \dots (n-1)}}{1 - b_{2 \cdot 34 \dots (n-1)} b_{n \cdot 34 \dots (n-1)}} \right] \dots (10.48a)$$

**10.18. Expression for Partial Correlation Coefficient in Terms of Correlation Coefficients of Lower Order.** By definition, we have

$$\begin{aligned}
 b_{ij \cdot lm \dots t} & = r_{ij \cdot lm \dots t} \times \frac{\sigma_{i \cdot lm \dots t}}{\sigma_{j \cdot lm \dots t}} \dots (*) \\
 \therefore b_{1 \cdot 34 \dots (n-1)} b_{n \cdot 34 \dots (n-1)} & \\
 & = r_{1 \cdot 34 \dots (n-1)} \frac{\sigma_{1 \cdot 34 \dots (n-1)}}{\sigma_{n \cdot 34 \dots (n-1)}} \times r_{n \cdot 34 \dots (n-1)} \frac{\sigma_{n \cdot 34 \dots (n-1)}}{\sigma_{2 \cdot 34 \dots (n-1)}} \\
 & = r_{1 \cdot 34 \dots (n-1)} \cdot r_{n \cdot 34 \dots (n-1)} \cdot \frac{\sigma_{1 \cdot 34 \dots (n-1)}}{\sigma_{2 \cdot 34 \dots (n-1)}} \dots (**)
 \end{aligned}$$

Hence from (10.48), on using (\*) and (\*\*), we get

$$\begin{aligned}
 r_{12 \cdot 34 \dots n} \times \frac{\sigma_{1 \cdot 34 \dots n}}{\sigma_{2 \cdot 34 \dots n}} \\
 = \left[ \frac{\{r_{12 \cdot 34 \dots (n-1)} - r_{1 \cdot 34 \dots (n-1)} r_{n \cdot 34 \dots (n-1)}\}}{1 - r_{2 \cdot 34 \dots (n-1)}^2} \frac{\sigma_{1 \cdot 34 \dots (n-1)}}{\sigma_{2 \cdot 34 \dots (n-1)}} \right] \dots (***)
 \end{aligned}$$

Also on using (10.47), we get

$$\frac{\sigma_{1 \cdot 34 \dots n}}{\sigma_{2 \cdot 34 \dots n}} = \frac{\sigma_{1 \cdot 34 \dots (n-1)}}{\sigma_{2 \cdot 34 \dots (n-1)}} \times \left[ \frac{1 - r_{1 \cdot 34 \dots (n-1)}^2}{1 - r_{2 \cdot 34 \dots (n-1)}^2} \right]^{1/2}$$

Hence from (\*\*), we get

$$\begin{aligned}
 r_{12 \cdot 34 \dots n} &= \left[ \frac{1 - r_{1n \cdot 34 \dots (n-1)}^2}{1 - r_{2n \cdot 34 \dots (n-1)}^2} \right]^{\frac{1}{2}} \\
 &= \left[ \frac{r_{12 \cdot 34 \dots (n-1)} - r_{1n \cdot 34 \dots (n-1)} r_{n2 \cdot 34 \dots (n-1)}}{1 - r_{2n \cdot 34 \dots (n-1)}^2} \right]^{\frac{1}{2}} \\
 \Rightarrow r_{12 \cdot 34 \dots n} &= \frac{r_{12 \cdot 34 \dots (n-1)} - r_{1n \cdot 34 \dots (n-1)} r_{n2 \cdot 34 \dots (n-1)}}{(1 - r_{1n \cdot 34 \dots (n-1)}^2)^{1/2} (1 - r_{n2 \cdot 34 \dots (n-1)}^2)^{1/2}} \quad \dots(10.49)
 \end{aligned}$$

which is an expression for the correlation coefficient of order  $p = (n-2)$  in terms of the correlation coefficient of order  $(p-1) = (n-3)$ .

**Example 10.33.** From the data relating to the yield of dry bark ( $X_1$ ), height ( $X_2$ ) and girth  $X_3$  for 18 cinchona plants the following correlation coefficients were obtained :

$$r_{12} = 0.77, \quad r_{13} = 0.72 \quad \text{and} \quad r_{23} = 0.52$$

Find the partial correlation coefficient  $r_{12 \cdot 3}$  and multiple correlation coefficient  $R_{1 \cdot 23}$ .

**Solution.**

$$\begin{aligned}
 r_{12 \cdot 3} &= \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \frac{0.77 - 0.72 \times 0.52}{\sqrt{[1 - (0.72)^2][1 - (0.52)^2]}} = 0.62 \\
 R_{1 \cdot 23}^2 &= \frac{r_{12}^2 + r_{13}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{23}^2} \\
 &= \frac{(0.77)^2 + (0.72)^2 - 2 \times 0.77 \times 0.72 \times 0.52}{1 - (0.52)^2} = 0.7334
 \end{aligned}$$

$$\therefore R_{1 \cdot 23} = \pm 0.8564$$

(since multiple correlation coefficient is non-negative).

**Example 10.34.** In a trivariate distribution :

$$\sigma_1 = 2, \sigma_2 = \sigma_3 = 3, \quad r_{12} = 0.7, \quad r_{23} = r_{31} = 0.5.$$

Find (i)  $r_{23 \cdot 1}$ , (ii)  $R_{1 \cdot 23}$ , (iii)  $b_{12 \cdot 3}$ ,  $b_{13 \cdot 2}$  and (iv)  $\sigma_{1 \cdot 23}$ .

**Solution.** We have

$$\begin{aligned}
 (i) \quad r_{23 \cdot 1} &= \frac{r_{23} - r_{21} r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} = \frac{0.5 - (0.7)(0.5)}{\sqrt{(1 - 0.49)(1 - 0.25)}} = 0.2425 \\
 (ii) \quad R_{1 \cdot 23}^2 &= \frac{r_{12}^2 + r_{13}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{23}^2} \\
 &= \frac{0.49 + 0.25 - 2(0.7)(0.5)(0.5)}{1 - 0.25} = 0.52
 \end{aligned}$$

$$\therefore R_{1 \cdot 23} = + 0.7211$$

$$(iii) \quad b_{12 \cdot 3} = r_{12 \cdot 3} \frac{\sigma_{1 \cdot 3}}{\sigma_{2 \cdot 3}} \quad \text{and} \quad b_{13 \cdot 2} = r_{13 \cdot 2} \frac{\sigma_{1 \cdot 2}}{\sigma_{3 \cdot 2}}$$

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = 0.6, \quad r_{13 \cdot 2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = 0.2425$$

$$\sigma_{1:3} = \sigma_1 \sqrt{(1 - r_{13}^2)} = 2 \sqrt{(1 - 0.25)} = 1.7320$$

$$\sigma_{2:3} = \sigma_2 \sqrt{(1 - r_{23}^2)} = 3 \sqrt{(1 - 0.25)} = 2.5980$$

$$\sigma_{1:2} = \sigma_1 \sqrt{(1 - r_{12}^2)} = 2 \sqrt{(1 - 0.49)} = 1.4282$$

$$\sigma_{3:2} = \sigma_3 \sqrt{(1 - r_{32}^2)} = 3 \sqrt{(1 - 0.25)} = 2.5980$$

Hence  $b_{12:3} = 0.4$  and  $b_{13:2} = 0.1333$

$$(iv) \quad \sigma_{1:23} = \sigma_1 \sqrt{\frac{\omega}{\omega_{11}}}$$

$$\text{where } \omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} = 0.36$$

$$\text{and } \omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - 0.25 = 0.75$$

$$\therefore \sigma_{1:23} = 2 \times \sqrt{0.48} = 2 \times 0.6928 = 1.3856$$

**Example 10.35.** Find the regression equation of  $X_1$  on  $X_2$  and  $X_3$  given the following results :—

Trait	Mean	Standard deviation	$r_{12}$	$r_{23}$	$r_{31}$
$X_1$	28.02	4.42	+ 0.80	—	—
$X_2$	4.91	1.10	—	-0.56	—
$X_3$	594	85	—	—	-0.40

where  $X_1$  = Seed per acre;  $X_2$  = Rainfall in inches

$X_3$  = Accumulated temperature above 42°F.

**Solution.** Regression equation of  $X_1$  on  $X_2$  and  $X_3$  is given by

$$(X_1 - \bar{X}_1) \frac{\omega_{11}}{\sigma_1} + (X_2 - \bar{X}_2) \frac{\omega_{12}}{\sigma_2} + (X_3 - \bar{X}_3) \frac{\omega_{13}}{\sigma_3} = 0$$

$$\text{where } \omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - (-0.56)^2 = 0.686$$

$$\omega_{12} = \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13}r_{23} - r_{21} = -0.576$$

$$\omega_{13} = r_{23}r_{12} - r_{13} = (-0.56)(0.80) - (-0.40) = -0.048$$

∴ Required equation of plane of regression of  $X_1$  on  $X_2$  and  $X_3$  is given by

$$\frac{0.686}{4.42} (X_1 - 28.02) + \frac{(-0.576)}{1.10} (X_2 - 4.91) + \frac{(-0.048)}{85.00} (X_3 - 594) = 0$$

**Example 10-36.** Five hundred students were examined in three subjects I, II and III, each subject carrying 100 marks. A student getting 120 or more but less than 150 marks was put in pass class. A student getting 150 or more but less than 180 marks was put in second class and a student getting 180 or more marks was put in the first class. The following marks were obtained :

	I	II	III
Mean :	35.8	52.4	48.8
S.D. :	4.2	5.3	6.1
Correlation :	$r_{12} = 0.6$ ,	$r_{13} = 0.7$	$r_{23} = 0.8$

- (i) Find the number of students in each of the three classes.  
(ii) Find the total number of students with total marks lying between 120 and 190.

(iii) Find the probability that a student gets more than 240 marks.

(iv) What should be the correlation between marks in subjects I and II among students who scored equal marks in subject III ?

(v) If  $r_{23}$  was not known, obtain the limits within which it may lie from the values of  $r_{12}$  and  $r_{13}$  (ignoring sampling errors).

**Solution.** If  $Z$  denotes the total marks of the students in the three subjects and  $X_1, X_2, X_3$  the total marks of the students in subjects I, II and III respectively, then

$$\begin{aligned} Z &= X_1 + X_2 + X_3 \\ \therefore E(Z) &= E(X_1) + E(X_2) + E(X_3) = 35.8 + 52.4 + 48.8 = 137 \\ V(Z) &= V(X_1) + V(X_2) + V(X_3) \\ &\quad + 2[\text{Cov}(X_1, X_2) + \text{Cov}(X_2, X_3) + \text{Cov}(X_3, X_1)] \\ &= 17.64 + 28.09 + 37.21 + 26.712 + 35.868 + 51.728 \\ &= 197.248 \quad [\text{Using } \text{Cov}(X_i, X_j) = r_{ij} \sigma_i \sigma_j] \\ \Rightarrow \sigma_Z^2 &= 197.248 \text{ or } \sigma_Z = 14.045 \\ \text{Now } \xi &= \frac{Z - E(Z)}{\sigma_Z} \sim N(0, 1) \end{aligned}$$

Z	$\xi = \frac{Z - 137}{14.045}$	$p = \int_{-\infty}^{\xi} p(\xi) d\xi$	Class	Area under the curve in this class (A)	Frequency $500 \times (A)$
120 -	1.21050	0.11314	120 - 150	0.70937	354.685
150	0.92567	0.82251	150 - 180	0.17639	88.195
180	3.06180	0.99890	180 -	0.00102	0.510
190	3.77400	0.99992	120 - 190	0.88678	443.390
240	7.33410	1.00000	240 -	0.00000	0.000

- (i) The number of students in first, second and third class respectively are 355, 88 and 0 (approx.).  
(ii) Total number of students with total marks between 120 and 190 is 443.  
(iii) Probability that a student gets more than 240 marks is zero.  
(iv) The correlation coefficient between marks in subjects I and II of the students who secured equal marks in subject III is  $r_{123}$  and is given by

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \frac{0.04}{\sqrt{(1 - 0.49)(1 - 0.64)}} = 0.0934$$

(v) We have .

$$\begin{aligned} r_{12 \cdot 3}^2 &= \frac{(r_{12} - r_{13} r_{23})^2}{(1 - r_{13}^2)(1 - r_{23}^2)} \leq 1 \\ \therefore \frac{(0.6 - 0.7a)^2}{(1 - 0.49)(1 - a^2)} &\leq 1, \text{ where } a = r_{23}. \\ \Rightarrow 0.36 + 0.49a^2 - 0.84a &\leq 0.51(1 - a^2) \\ \Rightarrow a^2 - 0.84a - 0.15 &\leq 0 \end{aligned}$$

Thus 'a' lies between the roots of the equation :

$$a^2 - 0.84a - 0.15 = 0,$$

which are 0.99 and -0.15.

Hence  $r_{23}$  should lie between -0.15 and 0.99.

**Example 10-37.** Show that

$$1 - R_{1 \cdot 23}^2 = (1 - r_{12}^2)(1 - r_{13 \cdot 2}^2)$$

Deduce that

$$(i) R_{1 \cdot 23} \geq r_{12}, \quad (ii) R_{1 \cdot 23}^2 = r_{12}^2 + r_{13}^2, \text{ if } r_{23} = 0$$

$$(iii) 1 - R_{1 \cdot 23}^2 = \frac{(1 - \rho)(1 + 2\rho)}{(1 + \rho)}, \text{ provided all the coefficients of zero order are equal to } \rho.$$

(iv) If  $R_{1 \cdot 23} = 0$ ,  $X_1$  is uncorrelated with any of other variables, i.e.,  $r_{12} = r_{13} = 0$ . [Delhi Univ. B.Sc. (Stat. Hons.), 1989]

**Solution.** (i) Since  $|r_{13 \cdot 2}| \leq 1$ , we have from (10-46c)

$$1 - R_{1 \cdot 23}^2 \leq 1 - r_{12}^2 \Rightarrow R_{1 \cdot 23} \geq r_{12}$$

(ii) We have

$$r_{13 \cdot 2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \frac{r_{13}}{\sqrt{1 - r_{12}^2}}. \quad (\text{if } r_{23} = 0)$$

∴ From (10-46c), we get

$$1 - R_{1 \cdot 23}^2 = (1 - r_{12}^2) \left[ 1 - \frac{r_{13}^2}{1 - r_{12}^2} \right] = 1 - r_{12}^2 - r_{13}^2$$

Hence  $R_{1 \cdot 23}^2 = r_{12}^2 + r_{13}^2$ , if  $r_{23} = 0$ .

(iii) Here, we are given that  $r_{12} = r_{13} = r_{23} = \rho$

$$\therefore r_{13 \cdot 2} = \frac{\rho - \rho^2}{\sqrt{(1 - \rho^2)(1 - \rho^2)}} = \frac{\rho(1 - \rho)}{(1 - \rho^2)} = \frac{\rho}{1 + \rho}$$

Hence from (10-46c), we have

$$1 - R_{1 \cdot 23}^2 = (1 - \rho^2) \left[ 1 - \frac{\rho^2}{(1 + \rho)^2} \right] = \frac{(1 - \rho)(1 + 2\rho)}{(1 + \rho)}$$

(iv) If  $R_{1 \cdot 23} = 0$ , (10-46c) gives

$$1 = (1 - r_{12}^2)(1 - r_{13 \cdot 2}^2)$$

...(\*)

Since  $0 \leq r_{12}^2 \leq 1$  and  $0 \leq r_{13-2}^2 \leq 1$ , (\*) will hold if and only if

$$r_{12} = 0 \quad \text{and} \quad r_{13-2} = 0$$

$$\text{Now } r_{13-2} = 0 \Rightarrow \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = 0$$

$$\Rightarrow \frac{r_{13}}{\sqrt{1 - r_{32}^2}} = 0 \quad (\because r_{12} = 0)$$

$$\Rightarrow r_{13} = 0$$

Thus if  $R_{1-23} = 0$ , then  $r_{13} = r_{12} = 0$ , i.e.,  $X_1$  is uncorrelated with  $X_2$  and  $X_3$ .

**Example 10-38.** Show that the correlation coefficient between the residuals  $X_{1-23}$  and  $X_{2-13}$  is equal and opposite to that between  $X_{1-3}$  and  $X_{2-3}$ .

[Poona Univ. B.Sc., 1991]

**Solution.** The correlation coefficient between  $X_{1-23}$  and  $X_{2-13}$  is given by

$$\begin{aligned} \frac{\text{Cov}(X_{1-23}, X_{2-13})}{\sigma_{1-23} \sigma_{2-13}} &= \frac{\sum X_{1-23} X_{2-13}}{N \sigma_{1-23} \sigma_{2-13}} = \frac{\frac{1}{N} \sum X_{2-13} (X_1 - b_{12-3} X_2 - b_{13-2} X_3)}{\sigma_{1-23} \sigma_{2-13}} \\ &= -b_{12-3} \frac{\sum X_{2-13} X_2}{N \sigma_{1-23} \sigma_{2-13}} \quad (\text{c.f. Property 1, § 10-13}) \\ &= -b_{12-3} \frac{\sum X_{2-13}^2}{N \sigma_{1-23} \sigma_{2-13}} \quad (\text{c.f. Property 2, § 10-13}) \\ &= -b_{12-3} \frac{\sigma_{2-13}}{\sigma_{1-23}} = -b_{12-3} \frac{(\sigma_2 \sqrt{\omega/\omega_{22}})}{(\sigma_1 \sqrt{\omega/\omega_{11}})} \end{aligned}$$

$$\text{where } \omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 \quad \text{and} \quad \omega_{22} = \begin{vmatrix} 1 & r_{13} \\ r_{31} & 1 \end{vmatrix} = 1 - r_{13}^2$$

$$\therefore r(X_{1-23}, X_{2-13}) = -b_{12-3} \frac{\sigma_2}{\sigma_1} \cdot \sqrt{\frac{1 - r_{23}^2}{1 - r_{13}^2}} = -b_{12-3} \frac{\sigma_{2-3}}{\sigma_{1-3}}$$

[since  $\sigma_{2-3}^2 = \sigma_2^2 (1 - r_{23}^2)$  and  $\sigma_{1-3}^2 = \sigma_1^2 (1 - r_{13}^2)$ ]

$$\therefore r(X_{1-23}, X_{2-13}) = -\frac{\text{Cov}(X_{1-3}, X_{2-3})}{\sigma_{2-3}^2} \cdot \frac{\sigma_{2-3}}{\sigma_{1-3}}$$

$$= -\frac{\text{Cov}(X_{1-3}, X_{2-3})}{\sigma_{2-3} \sigma_{1-3}} = -r(X_{1-3}, X_{2-3})$$

Hence the result.

**Example 10-39.** Show that if  $X_3 = aX_1 + bX_2$ , the three partial correlations are numerically equal to unity,  $r_{13-2}$  having the sign of  $a$ ,  $r_{23-1}$  the sign of  $b$  and  $r_{12-3}$  the opposite sign of  $a/b$ .

[Kanpur Univ. M.Sc., 1992]

**Solution.** Here we may regard  $X_3$  as dependent on  $X_1$  and  $X_2$  which may be taken as independent variables. Since  $X_1$  and  $X_2$  are independent, they are uncorrelated.

$$\text{Thus } r(X_1, X_2) = 0 \Rightarrow \text{Cov}(X_1, X_2) = 0$$

$$\begin{aligned} V(X_3) &= V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) + 2ab \text{Cov}(X_1, X_2) \\ &= a^2\sigma_1^2 + b^2\sigma_2^2. \end{aligned}$$

$$\text{where } V(X_1) = \sigma_1^2, V(X_2) = \sigma_2^2.$$

$$\text{Also } X_1 X_3 = X_1(aX_1 + bX_2) = aX_1^2 + bX_1 X_2$$

Assuming that  $X_i$ 's are measured from their means, on taking expectations of both sides, we get

$$\text{Cov}(X_1, X_3) = a\sigma_1^2 + b \text{Cov}(X_1, X_2) = a\sigma_1^2$$

$$\therefore r_{13} = \frac{\text{Cov}(X_1, X_3)}{\sqrt{V(X_1)V(X_3)}} = \frac{a\sigma_1^2}{\sqrt{\sigma_1^2(a^2\sigma_1^2 + b^2\sigma_2^2)}} = \frac{a\sigma_1}{k},$$

$$\text{where } k^2 = a^2\sigma_1^2 + b^2\sigma_2^2.$$

Similarly, we will get

$$r_{23} = \frac{\text{Cov}(X_2, X_3)}{\sqrt{V(X_2)V(X_3)}} = \frac{b\sigma_2}{k}$$

Hence

$$r_{13,2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \frac{a\sigma_1}{k} \frac{k}{\sqrt{k^2 - b^2\sigma_2^2}} = \frac{a\sigma_1}{\sqrt{a^2\sigma_1^2}} = \pm \frac{a\sigma_1}{a\sigma_1} = \pm 1$$

according as 'a' is positive or negative. Hence  $r_{13,2}$  has the same sign as 'a'.

Again

$$r_{23,1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} = \frac{b\sigma_2}{k} \frac{k}{\sqrt{k^2 - a^2\sigma_1^2}} = \frac{b\sigma_2}{\sqrt{b^2\sigma_2^2}} = \pm 1,$$

according as 'b' is positive or negative. Hence  $r_{23,1}$  has the same sign as 'b'.

Now

$$\begin{aligned} r_{12,3} &= \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = -\frac{a\sigma_1}{k} \cdot \frac{b\sigma_2}{k} \cdot \frac{k^2}{\sqrt{(k^2 - a^2\sigma_1^2)(k^2 - b^2\sigma_2^2)}} \\ &= -\frac{ab\sigma_1\sigma_2}{\sqrt{b^2\sigma_2^2 \times a^2\sigma_1^2}} = -\frac{ab}{\sqrt{a^2b^2}} = -\frac{a/b}{\sqrt{a^2/b^2}} = -\frac{(a/b)}{\pm(a/b)} = \mp 1, \end{aligned}$$

according as  $(a/b)$  is positive or negative. Hence  $r_{12,3}$  has the sign opposite to that of  $(a/b)$ .

**Example 10-40.** If all the correlation coefficients of zero order in a set of  $p$ -variates are equal to  $\rho$ , show that

(i) Every partial correlation of  $s$ 'th order is  $\frac{\rho}{1 + s\rho}$  ...(\*)

(ii) The coefficient of multiple correlation  $R$  of a variate with the other  $(p - 1)$  variates is given by

$$1 - R^2 = (1 - \rho) \left[ \frac{1 + (p - 1)\rho}{1 + (p - 2)\rho} \right]$$

**Solution.** We are given that

$$r_{mn} = \rho, (m, n = 1, 2, \dots, p; m \neq n)$$

We have

$$\begin{aligned} r_{ij\cdot k} &= \frac{r_{ij} - r_{ik} r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}, \quad (i, j, k = 1, 2, \dots, p; i \neq j \neq k) \\ &= \frac{\rho - \rho \cdot \rho}{\sqrt{(1 - \rho^2)(1 - \rho^2)}} = \frac{\rho}{1 + \rho} \end{aligned} \quad \dots (**)$$

Thus every partial correlation coefficient of first order is  $\rho/(1 + \rho)$ .

$\Rightarrow (*)$  is true for  $s = 1$ .

The result will be established by the principle of mathematical induction. Let us suppose that every partial correlation coefficient of order  $s$  is given by  $\rho/(1 + s\rho)$ . Then the partial correlation coefficient of order  $(s + 1)$  is given by

$$r_{ij\cdot km\ldots t} = \frac{r_{ij\cdot(s)} - r_{ik\cdot(s)} r_{jk\cdot(s)}}{\sqrt{(1 - r_{ik\cdot(s)}^2)(1 - r_{jk\cdot(s)}^2)}}$$

where  $k, m, \dots, t$  are  $(s + 1)$  secondary subscripts and  $r_{ij\cdot(s)}, r_{ik\cdot(s)}, r_{jk\cdot(s)}$ , are partial correlation coefficients of order  $s$ . Thus

$$r_{ij\cdot km\ldots t} = \frac{\frac{\rho}{1 + s\rho} - \left(\frac{\rho}{1 + s\rho}\right)^2}{1 - \left(\frac{\rho}{1 + s\rho}\right)^2} = \frac{\frac{\rho}{1 + s\rho} \left(1 - \frac{\rho}{1 + s\rho}\right)}{\left(1 - \frac{\rho}{1 + s\rho}\right)\left(1 + \frac{\rho}{1 + s\rho}\right)} \frac{\rho}{1 + (s + 1)\rho}$$

Using  $(**)$  and  $(***)$ , the required result follows by induction.

$$(ii) \text{ We have } 1 - R^2 = \frac{\omega}{\omega_{11}}$$

where  $R$  is the multiple correlation coefficient of a variable with other  $(p - 1)$  variables and

$$\omega = \begin{vmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{vmatrix}, \text{ a determinant of order 'p' and}$$

$$\omega_{11} = \begin{vmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{vmatrix} \text{ a determinant of order } (p - 1).$$

We have

$$\omega = [1 + (p - 1)\rho] \begin{vmatrix} 1 & \rho & \rho & \rho & \dots & \rho \\ 1 & 1 & \rho & \rho & \dots & \rho \\ 1 & \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \rho & \rho & \rho & \dots & 1 \end{vmatrix} \quad (\text{On adding } C_2, C_3, \dots, C_p \text{ to } C_1).$$

$$\Rightarrow \omega = [1 + (p - 1)\rho] \begin{vmatrix} 1 & \rho & \rho & \rho & \dots & \rho \\ 0 & (1 - \rho) & 0 & 0 & \dots & 0 \\ 0 & 0 & (1 - \rho) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & (1 - \rho) \end{vmatrix}$$

[On operating  $R_i - R_1$ , ( $i = 2, 3, \dots, p$ )].

$$\therefore \omega = [1 + (p - 1)\rho](1 - \rho)^{p-1}$$

Similarly, we will have

$$\omega_{11} = [1 + (p - 2)\rho](1 - \rho)^{p-2}.$$

$$\therefore 1 - R^2 = \frac{\omega}{\omega_{11}} = (1 - \rho) \left[ \frac{1 + (p - 1)\rho}{1 + (p - 2)\rho} \right]$$

**Example 10.41.** In a  $p$ -variate distribution, all the total (order zero) correlation coefficients are equal to  $\rho_0 \neq 0$ . Let  $\rho_k$  denote the partial correlation coefficient of order  $k$  and  $R_k$  be the multiple correlation coefficient of one variate on  $k$  other variates. Prove that

$$(i) \rho_0 \geq -\frac{1}{(p - 1)}, \quad (ii) \rho_k - \rho_{k-1} = -\rho_k \rho_{k-1}, \text{ and}$$

$$(iii) R_k^2 = \frac{k \rho_0^2}{1 + (k - 1)\rho_0}. \quad [\text{Delhi Univ. M.Sc. (Stat.) 1987}]$$

**Solution.** (i) We have proved in Example 10.40, that

$$\rho_k = \frac{\rho_0}{1 + k\rho_0}$$

In the case of  $p$ -variate distribution, the partial correlation coefficient of the highest order is  $\rho_{p-2}$  and is given by

$$\rho_{p-2} = \frac{\rho_0}{1 + (p - 2)\rho_0}$$

Since  $|\rho_{p-2}| \leq 1 \Rightarrow -1 \leq \rho_{p-2} \leq 1$ ,

we have (on considering the lower limit)

$$-1 \leq \frac{\rho_0}{1 + (p - 2)\rho_0} \quad \text{or} \quad -[1 + (p - 2)\rho_0] \leq \rho_0$$

$$\Rightarrow \rho_0 \geq -\frac{1}{(p - 1)}$$

$$\begin{aligned}
 (ii) \text{ L.H.S.} &= \rho_k - \rho_{k-1} = \frac{\rho_0}{1 + k\rho_0} - \frac{\rho_0}{1 + (k-1)\rho_0} \\
 &= \rho_0 \left[ \frac{-\rho_0}{(1 + k\rho_0)[1 + (k-1)\rho_0]} \right] \\
 &= - \left( \frac{\rho_0}{1 + k\rho_0} \right) \left( \frac{\rho_0}{1 + (k-1)\rho_0} \right) = -\rho_k \rho_{k-1}
 \end{aligned}$$

(iii) Taking  $\rho = \rho_0$  and  $k = p - 1$  in part (ii) Example 10.40, we get

$$\begin{aligned}
 1 - R_k^2 &= (1 - \rho_0) \left[ \frac{1 + k\rho_0}{1 + (k-1)\rho_0} \right] \\
 \therefore R_k^2 &= 1 - \frac{(1 - \rho_0)(1 + k\rho_0)}{1 + (k-1)\rho_0} = \frac{k \rho_0^2}{1 + (k-1)\rho_0} \quad (\text{On simplification})
 \end{aligned}$$

**Example 10.42.** If  $r_{12}$  and  $r_{13}$  are given, show that  $r_{23}$  must lie in the range :  $r_{12} r_{13} \pm (1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2)^{1/2}$

If  $r_{12} = k$  and  $r_{13} = -k$ , show that  $r_{23}$  will lie between  $-1$  and  $1 - 2k^2$ .

[Sardar Patel Univ. B.Sc. Oct., 1992; Madras Univ. B.Sc. (Stat. Main) 1991]

**Solution.** We have

$$\begin{aligned}
 r_{12,3}^2 &= \left[ \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \right]^2 \leq 1 \\
 \therefore (r_{12} - r_{13} r_{23})^2 &\leq (1 - r_{13}^2)(1 - r_{23}^2) \\
 \Rightarrow r_{12}^2 + r_{13}^2 r_{23}^2 - 2r_{12} r_{13} r_{23} &\leq 1 - r_{13}^2 - r_{23}^2 + r_{13}^2 r_{23}^2 \\
 \Rightarrow r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12} r_{13} r_{23} &\leq 1 \quad \dots(*)
 \end{aligned}$$

This condition holds for consistent values of  $r_{12}$ ,  $r_{13}$  and  $r_{23}$ . (\*) may be rewritten as :

$$r_{23}^2 - (2r_{12} r_{13}) r_{23} + (r_{12}^2 + r_{13}^2 - 1) \leq 0.$$

Hence, for given values of  $r_{12}$  and  $r_{13}$ ,  $r_{23}$  must lie between the roots of the quadratic (in  $r_{23}$ ) equation

$$r_{23}^2 - (2r_{12} r_{13}) r_{23} + (r_{12}^2 + r_{13}^2 - 1) = 0,$$

which are given by :

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

Hence

$$\begin{aligned}
 r_{12} r_{13} - \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2} &\leq r_{23} \leq r_{12} r_{13} \\
 &\quad + \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2} \quad \dots(**)
 \end{aligned}$$

In other words,  $r_{23}$  must lie in the range

$$r_{12} r_{13} \pm \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2}$$

In particular, if  $r_{12} = k$  and  $r_{13} = -k$ , we get from (\*\*)

$$-k^2 - \sqrt{1 - k^2 - k^2 + k^4} \leq r_{23} \leq -k^2 + \sqrt{1 - k^2 - k^2 + k^4}$$

$$\Rightarrow -k^2 - (1 - k^2) \leq r_{23} \leq -k^2 + (1 - k^2)$$

$$\therefore -1 \leq r_{23} \leq 1 - 2k^2$$

## EXERCISE 10(g)

1. (a) Explain partial correlation and multiple correlation.

(b) Explain the concepts of multiple and partial correlation coefficients.

Show that the multiple correlation coefficient  $R_{1.23}$  is, in the usual notations given by :

$$R_{1.23}^2 = 1 - \frac{\omega}{\omega_{11}}$$

2 (a) In the usual notations, prove that

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2} \leq r_{12}^2$$

(b) If  $R_{1.23} = 1$ , prove that  $r_{2.13}$  is also equal to 1. If  $R_{1.23} = 0$ , does it necessarily mean that  $R_{2.13}$  is also zero?

3. (a) Obtain an expression for the variance of the residual  $X_{1.23}$  in terms of the correlations  $r_{12}$ ,  $r_{23}$  and  $r_{31}$  and deduce that  $R_{1(23)} \geq r_{12}$  and  $r_{13}$ .

(b) Show that the standard deviation of order  $p$  may be expressed in terms of standard deviation of order  $(p - 1)$  and a correlation coefficient of order  $(p - 1)$ . Hence deduce that :

$$(i) \sigma_1 \geq \sigma_{1.2} \geq \sigma_{1.23} \geq \dots \geq \sigma_{1.23\dots n}$$

$$(ii) 1 - R_{1.23\dots n}^2 = (1 - r_{12}^2)(1 - r_{13}^2)\dots(1 - r_{1n-23\dots(n-1)}^2)$$

[Delhi Univ. M.Sc. (Stat.) 1987]

4. (a) In a  $p$ -variate distribution all the total (zero order) correlation coefficients are equal to  $\rho_0 \neq 0$ . If  $\rho_k$  denotes the partial correlation coefficient of order  $k$ , find  $\rho_k$ . Hence deduce that :

$$(i) \rho_k - \rho_{k-1} = -\rho_k \rho_{k-1}$$

$$\therefore (i) \rho_0 \geq -1/(p-1).$$

[Delhi Univ. M.Sc. (Stat.), 1989]

(b) Show that the multiple correlation coefficient  $R_{1.23\dots j}$  between  $X_1$  and  $(X_2, X_3, \dots, X_j)$ ,  $j = 2, 3, \dots, p$  satisfies the inequalities :

$$R_{1.2} \leq R_{1.23} \leq \dots \leq R_{1.23\dots p}$$

[Delhi Univ. M.Sc. (Maths.), 1989]

5. (a)  $X_0, X_1, \dots, X_n$  are  $(n+1)$  variates. Obtain a linear function of  $X_1, X_2, \dots, X_n$  which will have a maximum correlation with  $X_0$ . Show that the correlation  $R$  of  $X_0$  with the linear function is given by

$$R = \left(1 - \frac{\omega}{\omega_{00}^2}\right)^{\frac{1}{2}}$$

where  $\omega = \begin{vmatrix} 1 & r_{01} & r_{02}, \dots, r_{0n} \\ r_{10} & 1 & r_{12}, \dots, r_{1n} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ r_{n0} & r_{n1} & r_{n2}, \dots, 1 \end{vmatrix}$

and  $\omega_{00}$  is the determinant obtained by deleting the first row and the first column of  $\omega$ .

(b) With the usual notations, prove that

$$\sigma^2_{1,234,\dots,n} = \frac{\omega}{\omega_{11}} \sigma_1^2 = \sigma_1^2 (1 - r_{12}^2)(1 - r_{13,2}^2) \dots (1 - r_{1,n-23,\dots,n-1}^2)$$

(c) For a trivariate distribution, prove that

$$r_{12,3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

6. (a) The simple correlation coefficients between temperature ( $X_1$ ), corn yield ( $X_2$ ) and rainfall ( $X_3$ ) are,  $r_{12} = 0.59$ ,  $r_{13} = 0.46$  and  $r_{23} = 0.77$ .

Calculate the partial correlation coefficients  $r_{12,3}$ ,  $r_{23,1}$  and  $r_{31,2}$ . Also calculate  $R_{1,23}$ .

(b) If  $r_{12} = 0.80$ ,  $r_{13} = -0.40$  and  $r_{23} = -0.56$ , find the values of  $r_{12,3}$ ,  $r_{13,2}$  and  $r_{23,1}$ . Calculate further  $R_{1(23)}$ ,  $R_{2(13)}$  and  $R_{3(12)}$ .

7. (a) In certain investigation, the following values were obtained :

$$r_{12} = 0.6, r_{13} = -0.4 \text{ and } r_{23} = 0.7$$

Are the values consistent ?

(b) Comment on the consistency of

$$r_{12} = \frac{3}{5}, r_{23} = \frac{4}{5}, r_{31} = -\frac{1}{2}.$$

(c) Suppose a computer has found, for a given set of values of  $X_1$ ,  $X_2$  and  $X_3$ ,

$$r_{12} = 0.91, \quad r_{13} = 0.33 \text{ and } r_{23} = 0.81$$

Examine whether the computations may be said to be free from error.

8. (a) Show that if  $r_{12} = r_{13} = 0$ , then  $R_{1(23)} = 0$ . What is the significance of this result in regard to the multiple regression equation of  $X_1$  on  $X_2$  and  $X_3$  ?

(b) For what value of  $R_{1,23}$  will  $X_2$  and  $X_3$  be uncorrelated ?

(c) Given the data :  $r_{12} = 0.6$ ,  $r_{13} = 0.4$ , find the value of  $r_{23}$  so that  $R_{1,23}$ , the multiple correlation coefficient of  $X_1$  on  $X_2$  and  $X_3$  should be unity.

9. From the heights ( $X_1$ ), weights ( $X_2$ ) and ages ( $X_3$ ) of a group of students the following standard deviations and correlation coefficients were obtained :  $\sigma_1 = 2.8$  inches,  $\sigma_2 = 12$  lbs, and  $\sigma_3 = 1.5$  years,  $r_{12} = 0.75$ ,  $r_{23} = 0.54$ , and  $r_{31} = 0.43$ . Calculate (i) partial regression coefficients and (ii) partial correlation coefficients.

10. For a trivariate distribution :

$\bar{X}_1 = 40$	$\bar{X}_2 = 70$	$\bar{X}_3 = 90$
$\sigma_1 = 3$	$\sigma_2 = 6$	$\sigma_3 = 7$
$r_{12} = 0.4$	$r_{23} = 0.5$	$r_{13} = 0.6$

Find

(i)  $R_{1.23}$ , (ii)  $r_{23.1}$ , (iii) the value of  $X_3$  when  $X_1 = 30$  and  $X_2 = 45$ .

11. (a) In a study of a random sample of 120 students, the following results are obtained :

$$\begin{array}{lll} \bar{X}_1 = 68, & \bar{X}_2 = 70, & \bar{X}_3 = 74 \\ S_1^2 = 100, & S_2^2 = 25, & S_3^2 = 81, \\ r_{12} = 0.60, & r_{13} = 0.70, & r_{23} = 0.65 \end{array}$$

[ $S_i^2 = \text{Var}(X_i)$ ], where  $X_1$ ,  $X_2$ ,  $X_3$  denote percentage of marks obtained by a student in I test, II test and the final examination respectively.

(i) Obtain the least square regression equation of  $X_3$  on  $X_1$  and  $X_2$ .

(ii) Compute  $r_{12.3}$  and  $R_{3.12}$ .

(iii) Estimate the percentage marks of a student in the final examination if he gets 60% and 67% in I and II tests respectively.

(b)  $X_1$  is the consumption of milk per head,  $X_2$  the mean price of milk, and  $X_3$ , the per capita income. Time series of the three variables are rendered trend free and the standard deviations and correlation coefficients calculated :

$$s_1 = 7.22, \quad s_2 = 5.47, \quad s_3 = 6.87$$

$$r_{12} = -0.83, \quad r_{13} = 0.92, \quad r_{23} = -0.61$$

Calculate the regression equation of  $X_1$  on  $X_2$  and  $X_3$  and interpret the regression as a demand equation.

12. (a) Five thousand candidates were examined in the subjects (a), (b); (c); each of these subjects carrying 100 marks. The following constants relate to these data :

	<i>Subjects</i>		
	(a)	(b)	(c)
Mean	39.46	52.31	45.26
Standard deviation	6.2	9.4	8.7
$r_{bc} = 0.47$	$r_{ca} = 0.38$	$r_{ab} = 0.29$	

Assuming normally correlated population, find the number of candidates who will pass if minimum pass marks are an aggregate of 150 marks for the three subjects together.

(b) Establish the equation of plane of regression for variates  $X_1$ ,  $X_2$ ,  $X_3$  in the determinant form

$$\begin{vmatrix} X_1/\sigma_1 & X_2/\sigma_2 & X_3/\sigma_3 \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} = 0$$

[Delhi Univ. B.Sc. (Maths. Hons.), 1986]

13. (a) Prove the identity

$$b_{12.3} b_{23.1} b_{31.2} = r_{12.3} r_{23.1} r_{31.2}$$

[Gujarat Univ. B.Sc., 1992]

(b) Prove that

$$R_{1 \cdot 23}^2 = b_{12 \cdot 3} r_{12} \frac{\sigma_2}{\sigma_1} + b_{13 \cdot 2} r_{13} \frac{\sigma_3}{\sigma_1}$$

[Sardar Patel Univ. B.Sc., 1991]

14. (a) If  $X_3 = aX_1 + bX_2$  for all sets of values of  $X_1$ ,  $X_2$ , and  $X_3$ , find the value of  $r_{23 \cdot 1}$ .

(b) If the relation  $aX_1 + bX_2 + cX_3 = 0$  holds for all sets of values  $X_1$ ,  $X_2$  and  $X_3$ , what must be the partial correlation coefficients?

15. (a) If  $r_{12} = r_{23} = r_{31} = \rho \neq 1$ , then

$$r_{12 \cdot 3} = r_{23 \cdot 1} = r_{31 \cdot 2} = \frac{\rho}{1 + \rho} \text{ and } R_{1(23)} = R_{2(13)} = R_{3(12)} = \frac{\rho \sqrt{2}}{\sqrt{(1 + \rho)^2}}$$

(b)  $Y_1$ ,  $Y_2$ ,  $Y_3$  are uncorrelated standard variates.  $X_1 = Y_2 + Y_3$ ,  $X_2 = Y_3 + Y_1$ , and  $X_3 = Y_1 + Y_2$ . Find the multiple correlation coefficient between  $X_3$  and  $(X_1 \text{ and } X_2)$ .

16.  $X$ ,  $Y$ ,  $Z$  are independent random variables with the same variance. If

$$X_1 = \frac{1}{\sqrt{2}}(X - Z), X_2 = \frac{1}{\sqrt{3}}(X + Y + Z), X_3 = \frac{1}{\sqrt{6}}(X + 2Y + Z),$$

show that  $X_1$ ,  $X_2$ ,  $X_3$  have equal variances. Calculate  $r_{12 \cdot 3}$  and  $R_{1(23)}$ .

17. (a) If  $X_1$ ,  $X_2$  and  $X_3$  are three variables measured from their respective means as origin and if  $e_1$  is the expected value of  $X_1$  for given values of  $X_2$  and  $X_3$  from the linear regression of  $X_1$  on  $X_2$  and  $X_3$ , prove that

$$\text{Cov}(X_1, e_1) = \text{Var}(e_1) = \text{Var}(X_1) - \text{Var}(X_1 - e_1)$$

(b) If  $r_{12} = k$  and  $r_{23} = -k$ , show that  $r_{13}$  will lie between  $-1$  and  $1 - 2k^2$ .

18. (a) For three variables  $X$ ,  $Y$  and  $Z$ , prove that

$$r_{XY} + r_{YZ} + r_{ZX} \geq -\frac{3}{2} \quad \dots (*)$$

**Hint.** Let us transform  $X$ ,  $Y$ ,  $Z$  to their standard variables  $U$ ,  $V$  and  $W$ , (say), respectively, where

$$U = \frac{X - E(X)}{\sigma_X}, V = \frac{Y - E(Y)}{\sigma_Y}, W = \frac{Z - E(Z)}{\sigma_Z}$$

so that

$$E(U) = E(V) = E(W) = 0 \quad \left. \begin{array}{l} \sigma_U^2 = \sigma_V^2 = \sigma_W^2 = 1 \Rightarrow E(U^2) = E(V^2) = E(W^2) = 1 \\ \text{and} \end{array} \right\} \dots (**)$$

$$r_{UV} = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{E(UV) - E(U)E(V)}{\sigma_U \sigma_V} = E(UV) \quad \left. \begin{array}{l} \\ \text{and} \end{array} \right\} \dots (***)$$

Since correlation coefficient is independent of change of origin and scale, proving (\*) is equivalent to proving

$$r_{UV} + r_{VW} + r_{UW} \geq -3/2 \quad \dots (****)$$

To establish (\*\*\*\*) let us consider the  $E(U + V + W)^2$ , which is always non-negative i.e.,  $E(U + V + W)^2 \geq 0$ , and use (\*\*) and (\*\*\*).

(b)  $X, Y, Z$  are three reduced (standard) variates and  $E(YZ) = E(ZX) = -1/2$ , find the limits between which the coefficient of correlation  $r(X, Y)$  is necessarily placed.

**Hint.** Consider  $E(X + Y + Z)^2 \geq 0 \Rightarrow r \geq -\frac{1}{2}$ .

(c) If  $r_{12}, r_{23}$  and  $r_{31}$  are correlation coefficients of any three random variables  $X_1, X_2$  and  $X_3$  taken in pairs  $(X_1, X_2)$ ,  $(X_2, X_3)$  and  $(X_3, X_1)$  respectively, show that

$$1 + 2r_{12}r_{23}r_{31} \geq r_{12}^2 + r_{13}^2 + r_{23}^2$$

19. (a) If the relation  $aX_1 + bX_2 + cX_3 = 0$ , holds for all sets of values of  $X_1, X_2$  and  $X_3$ , where  $X_1, X_2$  and  $X_3$  are three standardised variables, find the three total correlation coefficients  $r_{12}, r_{23}$  and  $r_{13}$  in terms of  $a, b$  and  $c$ . What are the values of partial correlation coefficients if  $a, b$  and  $c$  are positive?

(b) Suppose  $X_1, X_2$  and  $X_3$  satisfy the relation  $a_1X_1 + a_2X_2 + a_3X_3 = k$ .

(i) Determine the three total correlation coefficients in terms of standard deviations and the constants  $a_1, a_2$  and  $a_3$ .

(ii) State what the partial correlation coefficients would be.

20. (a) Show that the multiple correlation between  $Y$  and  $X_1, X_2, \dots, X_p$  is the maximum correlation between  $Y$  and any linear function of  $X_1, X_2, \dots, X_p$ .

(b) Show that for  $p$  variates there are  ${}^pC_2$  correlation coefficients of order zero and  ${}^{p-2}C_s, {}^pC_2$  of order  $s$ . Show further that there are  ${}^pC_2, 2^{p-2}$  correlation coefficients altogether and  ${}^pC_2, 2^{p-1}$  regression coefficients.

### ADDITIONAL EXERCISES ON CHAPTER X

1. Find the correlation coefficient between

(i)  $aX + b$  and  $Y$ , (ii)  $lx + mY$  and  $X + Y$ , when correlation coefficient between  $X$  and  $Y$  is  $\rho$ .

2. If  $X_1$  and  $X_2$  are independent normal variates and  $U$  and  $V$  are defined by

$$U = X_1 \cos \alpha + X_2 \sin \alpha, \quad V = X_2 \cos \alpha - X_1 \sin \alpha,$$

show that the correlation coefficient  $\rho$  between  $U$  and  $V$  is given by

$$\rho^2 = 1 - \frac{4\sigma_1^2\sigma_2^2}{4\sigma_1^2\sigma_2^2 + (\sigma_1^2 - \sigma_2^2)\sin^2 2\alpha},$$

where  $\sigma_1^2$  and  $\sigma_2^2$  are variances of  $X_1$  and  $X_2$  respectively.

3. The variables  $X$  and  $Y$  are normally correlated, and  $\xi, \eta$  are defined by

$$\xi = X \cos \theta + Y \sin \theta, \quad \eta = Y \cos \theta - X \sin \theta$$

Obtain  $\theta$  so that the distributions of  $\xi$  and  $\eta$  are independent.

4. A set of  $n$  observations of simultaneous values of  $X$  and  $Y$  are made by an observer and the standard deviations and product moment coefficient about the mean are found to be  $\sigma_X, \sigma_Y$  and  $\rho_{XY}$ . A second observer repeating the same observations made a constant error  $e$  in observing each  $X$  and a constant error  $E$  in observing each  $Y$ . The two sets of observations are combined into a single set and coefficient of correlation calculated from it. Show that its value is

$$\sqrt{(\rho_{XY} + \frac{1}{4}eE) + \sqrt{(\sigma_X^2 + \frac{1}{4}e^2)(\sigma_Y^2 + \frac{1}{4}E^2)}}$$

**Hint.** here we have two sets of observations :

**1st Set :**  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ ; Mean =  $\bar{x}$ , s.d. =  $\sigma_x$ .

Product moment coefficient  $\rho_{xy} = r_{xy} \sigma_x \sigma_y$

**2nd Set :**  $(x_i + e, y_i + E)$ ,  $i = 1, 2, \dots, n$

$$\text{Mean } (\bar{x}') = \frac{1}{N} \sum (x_i + e) = \bar{x} + e$$

$$\text{Variance} = \sigma_x'^2 = \frac{1}{n} \sum [(x_i + e) - (\bar{x} + e)]^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \sigma_x^2$$

$$\text{Mean } (\bar{y}') = \bar{y} + E, \sigma_y'^2 = \sigma_y^2$$

Product moment coefficient :

$$\rho_{xy}' = \frac{1}{n} \sum [(x_i + e) - (\bar{x} + e)][(y_i + E) - (\bar{y} + E)] = \rho_{xy}$$

To obtain the correlation coefficient for the combined set of  $2n$  observations use Formula (10-5); Example 10-11(a) page 10-15.

5. Each of  $n$  independent trials can materialise in exactly one of the results

$A_1, A_2, \dots, A_k$ . If the probability of  $A_i$  is  $p_i$  in every trial  $\left( \sum_{i=1}^k p_i = 1 \right)$ , find the probability of obtaining the frequencies  $r_1, r_2, \dots, r_k$  for  $A_1, A_2, \dots, A_k$  respectively in these trials. Also find  $E(r_j)$ ,  $\text{Var}(r_j)$  and show that the correlation coefficient between  $r_i$  and  $r_j$  is independent of  $n$ .

6. In a sample of size  $n$  from a multinomial population  $n_1, n_2, \dots, n_k$  are of type  $1, 2, \dots, k$  with  $\sum p_i = 1$ , where  $p_i$  is the probability of type  $i$  ( $i = 1, 2, \dots, k$ ). Show that the expected value of  $n_2$  when  $n_1$  is given is  $(n - n_1) p_2 (1 - p_1)$  and hence or otherwise show that the coefficient of correlation between  $n_i$  and  $n_j$  is

$$= \left[ \frac{p_i p_j}{(1 - p_i)(1 - p_j)} \right]^{\frac{1}{2}}$$

7. A ball is drawn at random from an urn containing 3 white balls numbered 0, 1, 2; 2 red balls numbered 0, 1 and 1 black ball numbered 0. If the colours white, red and black are again numbered 0, 1 and 2 respectively, show that the correlation coefficient between the variables :  $X$ , the colour number and  $Y$ , the number of the ball is  $-\frac{1}{2}$ .

8. If  $X_1$  and  $X_2$  are two independent normal variates with a common mean zero and variances  $\sigma_1^2$  and  $\sigma_2^2$  respectively, show that the variates defined by

$$U_1 = X_1 + X_2 \quad \text{and} \quad U_2 = -\frac{\sigma_2}{\sigma_1} X_1 + \frac{\sigma_1}{\sigma_2} X_2$$

are independent and that each is normally distributed with mean zero and common variance  $(\sigma_1^2 + \sigma_2^2)$ .

9. If  $X_1, X_2$  and  $X_3$  are uncorrelated variables with equal mean  $M$  and variances  $V_1^2, V_2^2$  and  $V_3^2$  respectively, prove that correlation coefficient  $\rho$

between  $Z_1 = \frac{X_1}{X_3}$  and  $Z_2 = \frac{X_2}{X_3}$  is given by

$$\rho = \frac{V_3^2}{\sqrt{[(V_1^2 + V_3^2)(V_2^2 + V_3^2)]}}$$

**Hint.** Neglecting the cubes and higher powers of  $\frac{x_i}{M}$ ,  $x_i$  being the deviation of  $X_i$  from  $M$  and letting the means and s.d.'s of  $Z_1$  and  $Z_2$  to be  $I_1, I_2$  and  $s_1, s_2$  respectively, we get

$$\begin{aligned} I_1 &= \frac{1}{N} \sum \frac{X_1}{X_3} = \frac{1}{N} \sum_i (x_{1i} + M)(x_{3i} + M)^{-1} \\ &= \frac{1}{N} \sum \left( 1 + \frac{x_{1i}}{M} \right) \left( 1 + \frac{x_{3i}}{M} \right)^{-1} \\ &= \frac{1}{N} \sum \left[ \left( 1 - \frac{x_{3i}}{M} + \frac{x_{3i}^2}{M^2} - \dots \right) + \frac{x_{1i}}{M} - \frac{x_{1i}x_{3i}}{M^2} + \dots \right] \\ &= 1 + \frac{V_3^2}{M^2} \end{aligned}$$

$$\text{Similarly } I_2 = 1 + \frac{V_2^2}{M^2}$$

$$\therefore I_1 = I_2$$

$$\text{Now } s_1^2 = \frac{1}{N} \sum \left( \frac{X_1}{X_3} \right)^2 - I_1^2$$

$$\text{or } s_1^2 + I_1^2 = 1 + \frac{3V_3^2}{M^2} + \frac{V_1^2}{M^2}, \text{ and so we have } s_1^2 = \frac{V_3^2}{M^2} + \frac{V_1^2}{M^2}.$$

$$\text{Similarly } s_2^2 = \frac{V_2^2}{M^2} + \frac{V_3^2}{M^2}$$

$$\text{Now } N\rho s_1 s_2 = \sum \left( \frac{X_1}{X_3} - I_1 \right) \left( \frac{X_2}{X_3} - I_2 \right) = \frac{V_3^2}{M^2} \quad (\text{On simplification})$$

$$\text{Hence } \rho = \frac{N\rho s_1 s_2}{s_1 s_2} = \frac{V_3^2}{\sqrt{(V_3^2 + V_1^2)} \sqrt{(V_3^2 + V_2^2)}}$$

10. (*Weldon's Dice Problem*).  $n$  white dice and  $m$  red dice are shaken together and thrown on a table. The sum of the dots on the upper faces are noted. The red dice are then picked up and thrown again among the white dice left on the table. The sum of the dice on the upper faces is again noted. What is the correlation coefficient between the first and the second sums?

Ans.  $n/(n+m)$

11. Random variables  $X$  and  $Y$  have zero means and non-zero variances  $\sigma_X^2$  and  $\sigma_Y^2$ . If  $Z = Y - X$ , then find  $\sigma_Z^2$  and the correlation coefficient  $\rho(X, Z)$  of  $X$  and  $Z$  in terms of  $\sigma_X, \sigma_Y$  and the correlation coefficient  $r(X, Y)$  of  $X$  and  $Y$ .

For certain data  $Y = 1.2X$  and  $X = 0.6Y$ , are the regression lines. Compute  $r(X, Y)$  and  $\sigma_x/\sigma_y$ . Also compute  $\rho(X, Z)$ , if  $Z = Y - X$ .

[Calcutta Univ. B.Sc. (Maths. Honors.), 1984]

12. An item (say, a pen) from a production line can be acceptable, repairable or useless. Suppose a production is stable and let  $p, q, r$  ( $p + q + r = 1$ ), denote the probabilities for three possible conditions of an item. If the items are put into lots of 100 :

- Derive an expression for the probability function of  $(X, Y)$  where  $X$  and  $Y$  are the number of items in the lots that are respectively in the first two conditions.
- Derive the moment generating function of  $X$  and  $Y$ .
- Find the marginal distribution  $X$ .
- Find the conditional distribution of  $Y$  given  $X = 90$ .
- Obtain the regression function of  $Y$  on  $X$ .

[Delhi Univ. M.A. (Eco.), 1985]

13. If the regression of  $X_1$  on  $X_2, \dots, X_p$  is given by :

$$E(X_1 | X_2, \dots, X_p) = \alpha + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_p X_p$$

and

$$\begin{vmatrix} \sigma_{22} & \sigma_{23} & \dots & \sigma_{2p} \\ \sigma_{32} & \sigma_{33} & \dots & \sigma_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p2} & \sigma_{p3} & \dots & \sigma_{pp} \end{vmatrix} > 0, \quad \begin{cases} \sigma_{ii} = \text{variances} \\ \sigma_{ij} = \text{covariances} \end{cases}$$

then the constants  $\alpha, \beta_2, \dots, \beta_p$  are given by

$$\alpha = \mu_1 + \frac{R_{12}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_2} \cdot \mu_2 + \frac{R_{13}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_3} \cdot \mu_3 + \dots + \frac{R_{1p}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_p} \cdot \mu_p$$

and

$$\beta_j = -\frac{R_{1j}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_j}, \quad (j = 1, 2, \dots, p)$$

where  $R_{ij}$  is the cofactor of  $\rho_{ij}$  in the determinant ( $R$ ) of the correlation matrix

$$R = \begin{vmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & \rho_{pp} \end{vmatrix}$$

[Delhi Univ. M.Sc. (Stat.), 1988]

14. Let  $X_1$  and  $X_2$  be random variables with means 0 and variances 1 and correlation coefficient  $\rho$ . Show that :

$$E[\max(X_1^2, X_2^2)] \leq 1 + \sqrt{1 - \rho^2}$$

Using the above inequality, show that for random variables  $X_1$  and  $X_2$  with means  $\mu_1$  and  $\mu_2$ , variances  $\sigma_1^2$  and  $\sigma_2^2$  and correlation coefficient  $\rho$  and for any  $k > 0$ ,

$$P[|X_1 - \mu_1| \geq k\sigma_1 \text{ or } |X_2 - \mu_2| \geq k\sigma_2] \leq \frac{1}{k^2} [1 + \sqrt{1 - \rho^2}]$$

15. Let the maximum correlation between  $X_0$  and any linear function of  $X_1, X_2, \dots, X_n$  be  $R$  and if  $r_{01} = r_{02} = \dots = r_{0n} = r$

and all other correlation coefficients are equal to  $s$ , then show that :

$$R = r \left[ \frac{n}{1 + (n - 1)s} \right]^{1/2}$$

16. If  $f = f(x, y)$  is the p.d.f. of  $BVN(0, 0, 1, 1, \rho)$  distribution, verify that :

$$\frac{\partial f}{\partial \rho} = \frac{\partial^2 f}{\partial x \partial y}$$

Further, if two new random variables  $U$  and  $V$  are defined by the relation

$$U = P(Z \leq x) \text{ and } V = P(Z \leq y) \text{ where } Z \sim N(0, 1),$$

prove the marginal distributions of both  $U$  and  $V$  are uniform in the interval  $(-\frac{1}{2}, \frac{1}{2})$  and their common variance is  $\frac{1}{12}$ .

Hence prove that  $R = \text{Corr.}(U, V)$ , satisfies the relation :  $\rho = 2 \sin(\pi R/6)$ .

[Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1988]

17. If  $(X, Y) \sim BVN(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , then prove that  $a + bX + cY$ , ( $b \neq 0, c \neq 0$ ) is distributed as  $N(a + b\mu_x + c\mu_y, b^2\sigma_x^2 + c^2\sigma_y^2 + 2bc\rho\sigma_x\sigma_y)$ .

[Delhi Univ. M.Sc. (Stat.), 1989]

18. Let  $X_1, X_2, X_3$  be a random sample of size  $n = 3$  from  $N(0, 1)$  distribution.

(a) Show that  $Y_1 = X_1 + \delta X_3, Y_2 = X_2 + \delta X_3$  has a bivariate normal distribution.

(b) Find the value of  $\delta$  so that  $\rho(Y_1, Y_2) = \frac{1}{2}$ .

(c) What additional transformation involving  $Y_1$  and  $Y_2$  would produce a bivariate normal distribution with means  $\mu_1$  and  $\mu_2$ , variances  $\sigma_1^2$  and  $\sigma_2^2$ , and the same correlation coefficient  $\rho$  ?

Ans. (b) -1 or 1. (c)  $Z_1 = \sigma_1 Y_1 + \mu_1, Z_2 = \sigma_2 Y_2 + \mu_2$ .

19. If  $(X, Y) \sim BVN(0, 0, 1, 1, \rho)$ , prove that :

$$E[\max(X, Y)] = [(1 - \rho)/\pi]^{1/2} \text{ and } E[\min(X, Y)] = -[(1 - \rho)/\pi]^{1/2}$$

20. If  $(X, Y) \sim BVN(0, 0, \sigma_1^2, \sigma_2^2, \rho)$ , show that  $r$ th cumulant of  $XY$  is given by :

$$\kappa_r = \frac{1}{2}(r - 1)! \sigma_1^r \sigma_2^r [(\rho + 1)^r + (\rho - 1)^r].$$

$$\text{Deduce that } E(X^2 Y^2) = \sigma_1^2 \sigma_2^2 (1 + 2\rho^2).$$

21. Let  $f$  and  $g$  be the p.d.f.'s of  $X$  and  $Y$  with corresponding distribution functions  $F$  and  $G$ . Also let

$$h(x, y) = f(x) g(y) [1 + \alpha (2f(x) - 1)(2G(x) - 1)]; |\alpha| \leq 1,$$

Show that  $h(x, y)$  is a joint p.d.f. with marginal p.d.f.'s  $f$  and  $g$ . Further, let  $f$  and  $g$  be  $N(0, 1)$  p.d.f.'s. Show that  $Z = X + Y$ , is not normally distributed, except in the trivial case  $\alpha = 0$ .

**Hint.** Find  $M_Z(t) = E(e^{tZ})$  and use  $\text{Cov}(X, Y) = \alpha/\pi$ .

22. State p.d.f. of bivariate normal distribution. Let  $X$  and  $Y$  have joint p.d.f. of the form :

$$f(x, y) = ke^{-\frac{1}{2} [a_{11}(x - b_1)^2 + 2a_{12}(x - b_1)(y - b_2) + a_{22}(y - b_2)^2]}; \\ -\infty < (x, y) <$$

Find (i)  $k$ , (ii) the correlation coefficient between  $X$  and  $Y$ .

23. Write down, but do not derive, the moment generating function for a pair of random variables which have a bivariate normal distribution with both means equal to zero.

The independent random variables  $X, Y, Z$ , are each normally distributed with mean 0 and variance 1. If  $U = X + Y + Z$  and  $V = X - Y + 2Z$ , show that  $U$  and  $V$  have bivariate normal distribution. Find the correlation of  $U$  with  $V$  and the expectation of  $U$  when  $V$  is equal to 1.

24. Let  $X_1$  and  $X_2$  have a joint m.g.f.

$$M(t_1, t_2) = [a(e^{t_1} + t_2 + 1) + b(e^{t_1} + e^{t_2})]^2$$

in which  $a$  and  $b$  are positive constants such that  $2a + 2b = 1$ .

Find  $E(X_1)$ ,  $E(X_2)$ ,  $\text{Var}(X_1)$ ,  $\text{Var}(X_2)$ ,  $\text{Cov}(X_1, X_2)$ .

Ans. Means = 1, Variances =  $\frac{1}{2}$ , Covariance =  $2a - \frac{1}{2}$ .

25.  $X_1, X_2, X_3$  have joint distribution as a multinomial distribution with parameters  $N, p_1, p_2, p_3$ . If  $r_{ij}$  is the correlation coefficient between  $X_i$  and  $X_j$ , find the expression for  $r_{12}, r_{23}$  and  $r_{31}$  and hence deduce the expression for the partial correlation coefficient  $r_{123}$ .

26. (i) If all the inter-correlations between  $(p+1)$  variates  $X_0, X_1, X_2, \dots, X_p$  are equal to  $r$ , show that each of the partial correlation co-efficients of order  $p-1$  is equal to  $r/[1+(p-1)r]$  and that the multiple correlation of  $X_0$  on  $X_1, X_2, \dots, X_p$  is given by

$$1 - R_{0(12\dots p)}^2 = \frac{(1-r)(1-pr)}{1+(p-1)r}$$

$$(ii) \quad r_{12} = (r_{123} - r_{13}r_{23})/\sqrt{(1-r_{13}^2)^{1/2}(1-r_{23}^2)^{1/2}}$$

- 27. If  $R$  denotes the multiple correlation co-efficient of  $X_1$  on  $X_2, X_3, \dots, X_p$  in  $p$ -variate distribution, prove that

(i)  $R^2 \geq R_0^2$ , where  $R_0$  is the correlation of  $X_1$  with any arbitrary linear function of  $X_2, X_3, \dots, X_p$ .

(ii)  $R^2 \geq R_1^2$ , where  $R_1$  is the multiple correlation coefficient of  $X_1$  with  $X_2, X_3, \dots, X_k, k < p$

$$(iii) \quad 1 - R^2 = \prod_{j=2}^p (1 - r_{1j;23\dots(j-1)}^2)$$

## Theory of Attributes

---

**11·1. Introduction.** Literally, an attribute means a *quality* or *characteristic*. Theory of attributes deals with qualitative characteristics which are not amenable to quantitative measurements and hence need slightly different statistical treatment from that of the variables. Examples of attributes are drinking, smoking, blindness, health, honesty, etc. An attribute may be marked by its presence (possession) or absence (dispossession) in a member of given population. It may be pointed out that the method of statistical analysis applicable to the study of variables can also be used to a great extent in the theory of attributes and *vice-versa*. For example, the presence or absence of an attribute may be regarded as changes in the values of a variable which can possess only two values *viz.* 0 and 1. For the sake of clarity and simplicity, the theory of attributes has been developed independently.

**11·2. Notations.** Suppose the population is divided into two classes, according to the presence or absence of a single attribute. The *positive class*, which denotes the presence of the attribute is generally written in capital Roman letters such as  $A, B, C, D$  etc. and the *negative class*, denoting the absence of the attribute is written in corresponding small Greek letters such as  $\alpha, \beta, \gamma, \delta$ , etc. For example if  $A$  represents the attribute sickness and  $B$  represents blindness, than  $\alpha$  and  $\beta$  represent the attributes non-sickness (health) and sight respectively. The two classes *viz.*,  $A$  (possession of the attribute) and  $\alpha$  (dispossession of the attribute) are said to be *complementary classes* and the attribute  $\alpha$  used in the sense of not- $A$  is often called the complementary attribute of  $A$ . Similarly  $\beta, \gamma, \delta$  are the complementary attributes of  $B, C, D$  respectively.

The combinations of attributes are denoted by grouping together the letters concerned *e.g.*  $AB$  is the combination of the attributes  $A$  and  $B$ . Thus for the attributes  $A$  (sickness) and  $B$  (smoking),  $AB$  would mean the simultaneous possession of sickness and smoking. Similarly  $A\beta$  will represent sickness and non-smoking,  $\alpha B$  non-sickness (health) and smoking, and  $\alpha\beta$  non-sickness and non-smoking.

If a third attribute be included to represent, say male, then  $ABC$  will stand for sick males who are smokers. Similar interpretations can be given to  $AB\gamma$ ,  $A\beta C$ ,  $A\beta\gamma$  etc.

**11·3. Dichotomy.** If the universe (population) is divided into two subclasses or complementary classes and no more, with respect to each of the attributes  $A, B, C$  etc., the division or classification is said to be '*dichotomous classification*'. The classification is termed *manifold* if each class is further subdivided.

**11·4. Classes and Class Frequencies.** Different attributes in themselves are called different classes and the number of observations assigned to

them are called *class frequencies* which are denoted by bracketing the class-symbols. Thus  $(A)$  stands for the frequency of  $A$  and  $(AB)$  for the number of objects possessing the attribute  $AB$ .

**Remark.** Class frequencies of the type  $(A)$ ,  $(AB)$ ,  $(ABC)$  etc. are known as *positive frequencies*;  $(\alpha)$ ,  $(\alpha\beta)$ ,  $(\alpha\beta\gamma)$  etc. are known as *negative frequencies* and  $(\alpha B)$ ,  $(A\beta C)$ ,  $(\alpha\beta C)$  etc. are called the *contrary frequencies*.

**11.4.1. Order of Classes and Class frequencies.** A class represented by  $n$  attributes is called a class of  $n$ th order and the corresponding frequency as the frequency of the  $n$ th order. Thus  $(A)$  is a class frequency of order 1;  $(AB)$ ,  $(AC)$ ,  $(B\gamma)$  etc. are class frequencies of second order;  $(ABC)$ ,  $(A\beta\gamma)$ ,  $(\alpha\beta C)$  etc. are frequencies of third order and so on.  $N$ , the total number of members of the population, without any specification of attributes, is reckoned as a frequency of zero-order.

Thus in a dichotomous classification with respect to  $n$  attributes, the number of class frequencies of order ' $r$ ' is  $\binom{n}{r} \cdot 2^r$ , since  $r$  attributes out of  $n$  can be selected in  $\binom{n}{r}$  ways and each of the  $r$  attributes contributes two symbols, one representing the positive part (e.g.,  $A$ ) and the other the negative part (e.g.,  $\alpha$ ). Thus the total number of class frequencies of all orders, for  $n$  attributes is :

$$\sum_{r=0}^n \binom{n}{r} 2^r = 1 + \binom{n}{1} 2 + \binom{n}{2} 2^2 + \dots + \binom{n}{n} 2^n = (1+2)^n = 3^n \quad \dots(11.1)$$

**Remarks 1.** In particular, for  $n$  attributes, the total number of class frequencies of different orders are given as follows :

Order	0	1	2	...	$r$	...	$n$
No. of frequencies	1	$2^n$	$\binom{n}{2} 2^2$	...	$\binom{n}{r} 2^r$	...	$2^n$

2. Since in the case of  $n$  attributes, the positive class frequency of order  $r$  has  $\binom{n}{r}$  elements, their total number is :

$$\sum_{r=0}^n \binom{n}{r} = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = (1+1)^n = 2^n$$

3. In case of 3 attributes  $A$ ,  $B$  and  $C$ , the total number of class frequencies is  $3^3 = 27$ , as given below :

Order	Frequencies			
0	$N$			
1	$(A)$ $(\alpha)$	$(B)$ $(\beta)$	$(C)$ $(\gamma)$	
2	$(AB)$ $(AC)$ $(BC)$	$(A\beta)$ $(A\gamma)$ $(B\gamma)$	$(\alpha B)$ $(\alpha C)$ $(\beta C)$	$(\alpha\beta)$ $(\alpha\gamma)$ $(\beta\gamma)$
3	$(ABC)$ $(\alpha BC)$	$(AB\gamma)$ $(\alpha B\gamma)$	$(A\beta C)$ $(\alpha\beta C)$	$(A\beta\gamma)$ $(\alpha\beta\gamma)$

...(11.2)

**11.4.2. Relation Between Class Frequencies.** All the class frequencies of various orders are not independent of each other and any class frequency can always be expressed in terms of class frequencies of higher order. Thus

$$N = (A) + (\alpha) = (B) + (\beta) = (C) + (\gamma), \text{ etc.}$$

Also, since each of these  $A$ 's or  $\alpha$ 's can either be  $B$ 's or  $\beta$ 's, we have

$$(A) = (AB) + (A\beta) \quad \text{and} \quad (\alpha) = (\alpha B) + (\alpha\beta)$$

$$\text{Similarly } (B) = (AB) + (\alpha B) \quad \text{and} \quad (\beta) = (A\beta) + (\alpha\beta)$$

$$(AB) = (ABC) + (AB\gamma), \quad (A\beta) = (A\beta C) + (A\beta\gamma)$$

$$(\alpha B) = (\alpha BC) + (\alpha B\gamma), \quad (\alpha\beta) = (\alpha\beta C) + (\alpha\beta\gamma)$$

and so on. Thus

$$(A) = (AB) + (A\beta) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$$

$$(\beta) = (A\beta) + (\alpha\beta) = (A\beta C) + (A\beta\gamma) + (\alpha\beta C) + (\alpha\beta\gamma), \text{ etc.}$$

The classes of highest order are called the *ultimate classes* and their frequencies, the *ultimate class frequencies*. Thus in case of  $n$  attributes, the ultimate class frequencies will be the frequencies of  $n$ th order. For example, the class frequencies  $(ABC)$ ,  $(AB\gamma)$ ,  $(A\beta C)$ ,  $(A\beta\gamma)$ ,  $(\alpha BC)$ ,  $(\alpha B\gamma)$ ,  $(\alpha\beta C)$ ,  $(\alpha\beta\gamma)$  are the ultimate frequencies for three attributes  $A$ ,  $B$  and  $C$ .

**Remarks** 1. In case of  $n$  attributes, the ultimate class frequencies each contain  $n$  symbols and since each symbol may be written in two ways, viz., positive part and negative part, e.g.,  $A$  or  $\alpha$ ,  $B$  or  $\beta$ , etc., the total number of ultimate class frequencies is  $2^n$ .

2. By expressing any class frequency in terms of the class frequency of higher order, we can express it ultimately as the sum of some of the  $2^n$  ultimate class frequencies.

3. The total number of ultimate class frequencies specify the data completely.

4. The set of ultimate class frequencies is not the only set which specify the data completely. In fact any set of class frequencies which are (i)  $2^n$  in number and (ii) which are algebraically independent of each other, will specify the data completely. Such a set is called the *Fundamental Set*. For example, the positive class frequencies form such a set. Thus for  $n = 2$ , the set of positive class frequencies  $2^2 = 4$  (c.f. 11.2), is  $N$ ,  $(A)$ ,  $(B)$ ,  $(AB)$ . If we are given these

frequencies, then it is obvious from the table that the remaining frequencies, viz.,  $(A\beta)$ ,  $(\alpha\beta)$ ,  $(\alpha B)$ ,  $(\alpha)$  and  $(\beta)$  can be obtained by subtraction, e.g., given :

	$A$	$\alpha$	Total
$B$	$(AB)$	-	$(B)$
$\beta$	-	-	$(\beta)$
Total	$(A)$	$(\alpha)$	$N$

$$(\alpha) = N - (A), (\beta) = N - (B)$$

$$(A\beta) = (A) - (AB), (\alpha B) = (B) - (AB)$$

$$(\alpha\beta) = (\alpha) - (\alpha B) = N - (A) - (B) + (AB)$$

### 11.5. Class Symbols as Operators.

Let us write symbolically

$$A.N = (A) \quad \dots(11.3)$$

which means that the operation of dichotomising  $N$  according to  $A$  gives the class frequency equal to  $(A)$ . Similarly, we write

$$\alpha.N = (\alpha)$$

Adding, we get

$$A.N + \alpha.N = (A) + (\alpha)$$

$$\Rightarrow (A + \alpha).N = (A) + (\alpha)$$

$$\Rightarrow (A + \alpha).N = N$$

$$\Rightarrow A + \alpha = 1$$

Thus in symbolic expression we can replace  $A$  by  $(1 - \alpha)$  and  $\alpha$  by  $(1 - A)$ . Similarly,  $B$  can be replaced by  $(1 - \beta)$  and  $\beta$  by  $(1 - B)$ , and so on.

Dichotomising  $(B)$  according to  $A$ , let us write

$$A.(B) = (AB)$$

Similarly,

$$B.(A) = (BA)$$

$$A.(B) = B.(A) = (AB) = AB.N,$$

which amounts to dichotomising  $N$  according to  $AB$ .

For example :

$$(\alpha\beta) = \alpha\beta.N = (1 - A)(1 - B).N = N - A.N - B.N + AB.N$$

$$= N - [(A) + (B)] + (AB)$$

$$(\alpha\beta\gamma) = \alpha\beta\gamma.N = (1 - A)(1 - B)(1 - C).N$$

$$= N - A.N - B.N - C.N + AB.N + AC.N + BC.N - ABC.N$$

$$= N - [(A) + (B) + (C)] + [(AB) + (AC) + (BC)] - (ABC)$$

$$(AB\gamma) = AB\gamma.N = AB(1 - C).N = AB.N - ABC.N$$

$$= (AB) - (ABC)$$

$$(\alpha\beta C) = (1 - A)(1 - B)C.N = (C - AC - BC + ABC).N$$

$$= (C) - (AC) - (BC) + (ABC)$$

and so on.

**Example 11.1.** An investigation of 23,713 households was made in an urban and rural mixed locality. Of these 1,618 were farmers, 2,015 well-to-do and 770 families were having at least one graduate. Of these graduate families 335 were those of farmers and 428 were well-to-do, also 587 well-to-do families were those of farmers and out of them only 156 were having at least one of their family member as graduate. Obtain all the ultimate class frequencies.

**Solution.** Let the attribute 'farming' be denoted by  $A$ , the attribute 'well-to-do' by  $B$  and 'having at least one graduate' by  $C$ . Then in the usual notations, we are given

$$N = 23713, \quad (A) = 1618, \quad (B) = 2015, \quad (C) = 770, \quad (AB) = 587, \\ (BC) = 428, \quad (AC) = 335 \text{ and } (ABC) = 156.$$

For three attributes  $A, B, C$ , the number of ultimate class frequencies is  $2^3 = 8$ , one of them being  $(ABC) = 156$ . The remaining frequencies are obtained below :

$$(AB\gamma) = (AB) - (ABC) = 587 - 156 = 431 \\ (A\beta C) = (AC) - (ABC) = 335 - 156 = 179 \\ (A\beta\gamma) = (A) - (AB) - (AC) + (ABC) \\ = 1618 - 587 - 335 + 156 = 852 \\ (\alpha BC) = (BC) - (ABC) = 428 - 156 = 272 \\ (\alpha B\gamma) = (B) - (AB) - (BC) + (ABC) \\ = 2015 - 587 - 428 + 156 = 1156 \\ (\alpha\beta C) = (C) - (AC) - (BC) + (ABC) = 770 - 335 - 428 + 156 = 163 \\ (\alpha\beta\gamma) = N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC) \\ = 23713 - 1618 - 2015 - 770 + 587 + 335 + 428 - 156 = 20504$$

**Example 11.2. (a)** Given the following ultimate class frequencies, find the frequencies of positive class,

$$(ABC) = 149, \quad (AB\gamma) = 738, \quad (A\beta C) = 225, \quad (A\beta\gamma) = 1,196 \\ (\alpha BC) = 204, \quad (\alpha B\gamma) = 1,762, \quad (\alpha\beta C) = 171 \quad \text{and} \quad (\alpha\beta\gamma) = 21,842$$

**(b)** Find the remaining class frequencies, given the following data :

$$N = 23,713, \quad (A) = 1618, \quad (B) = 2015, \quad (C) = 770 \\ (AB) = 587, \quad (AC) = 428, \quad (BC) = 335, \quad (ABC) = 156$$

**Solution.** (a)  $(A) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) = 2,308$

$$(B) = (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma) = 2,853$$

$$(C) = (ABC) + (A\beta C) + (\alpha BC) + (\alpha\beta C) = 749$$

$$(AB) = (ABC) + (AB\gamma) = 887$$

$$(AC) = (ABC) + (A\beta C) = 374$$

$$(BC) = (ABC) + (\alpha BC) = 353$$

and  $N = [(ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) + (\alpha BC) + (\alpha B\gamma) \\ + (\alpha\beta C) + (\alpha\beta\gamma)] = 26,287$

(b) For three attributes, there are  $3^3 = 27$ , class frequencies in all. Thus we have to determine the remaining 19 class frequencies :

**Order 1 :**

$$(\alpha) = N - (A) = 22,095; (\beta) = N - (B) = 21,698$$

$$(\gamma) = N - (C) = 22,943$$

*Order 2:*

$$\begin{aligned} (A\beta) &= (A) - (AB) = 1,031 \\ (\alpha B) &= (B) - (AB) = 1,428 \\ (\alpha\beta) &= (\alpha) - (\alpha B) = 20,667 \\ (A\gamma) &= (A) - (AC) = 1,190 \\ (\alpha C) &= (C) - (AC) = 342 \\ (\alpha\gamma) &= (\alpha) - (\alpha C) = 21,753 \\ (B\gamma) &= (B) - (BC) = 1,680 \\ (\beta C) &= (C) - (BC) = 435 \\ (\beta\gamma) &= (\beta) - (\beta C) = 21,263 \end{aligned}$$

*Order 3:*

$$\begin{aligned} (AB\gamma) &= (AB) - (ABC) = 431 \\ (A\beta C) &= (AC) - (ABC) = 272 \\ (A\beta\gamma) &= (A\beta) - (A\beta C) = 759 \\ (\alpha BC) &= (BC) - (ABC) = 179 \\ (\alpha B\gamma) &= (\alpha B) - (\alpha BC) = 1249 \\ (\alpha\beta C) &= (\beta C) - (A\beta C) = 163 \\ (\alpha\beta\gamma) &= (\alpha\beta) - (\alpha\beta C) = 20,504 \end{aligned}$$

**Example 11.3.** Show that for  $n$  attributes  $A_1, A_2, A_3, \dots, A_n$ ,

$$(A_1 A_2 A_3 \dots A_n) \geq (A_1) + (A_2) + (A_3) + \dots + (A_n) - (n-1)N \quad \dots(11.4)$$

where  $N$  is the total number of observations.

**Solution.** We have

$$(\alpha_1 \alpha_2) = \alpha_1 \alpha_2, N = (1 - A_1)(1 - A_2), N = N - (A_1) - (A_2) + (A_1 A_2)$$

Since class frequency is always non-negative, we have

$$(\alpha_1 \alpha_2) \geq 0 \Rightarrow (A_1 A_2) \geq (A_1) + (A_2) - N \quad \dots(*)$$

It follows that (11.4) is true for 2 attributes.

Let us now suppose that (11.4) is true for  $r$  attributes  $A_1, A_2, \dots, A_r$ , so that

$$(A_1 A_2 A_3 \dots A_r) \geq (A_1) + (A_2) + (A_3) + \dots + (A_r) - (r-1)N$$

Replacing the attribute  $A_r$  by another compound attribute  $A_r A_{r+1}$ , we get

$$\begin{aligned} (A_1 A_2 A_3 \dots A_r A_{r+1}) &\geq (A_1) + (A_2) + (A_3) + \dots + (A_r A_{r+1}) - (r-1)N \\ &\geq (A_1) + (A_2) + (A_3) + \dots + ((A_r) + (A_{r+1}) - N) - (r-1)N \\ &\quad [\text{From } (*)] \\ &= (A_1) + (A_2) + \dots + (A_r) + (A_{r+1}) - rN \end{aligned}$$

This implies that if (11.4) is true for  $n = r$ , it is also true for  $n = r + 1$  attributes. But we have seen in (\*) that (11.4) is true for  $n = 2$ . Hence by mathematical induction, the result is true for all positive integral values of  $n$ .

**Example 11.4.** Show that if  $A$  occurs in a larger proportion of the cases where  $B$  is than where  $B$  is not, then  $B$  will occur in a larger proportion of cases where  $A$  is than where  $A$  is not.

**Solution.** The problems can be restated as follows :

$$\text{Given } \frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)}, \text{ prove that } \frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$$

$$\text{Now } \frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)} \Rightarrow \frac{(\beta)}{(B)} > \frac{(A\beta)}{(AB)}$$

$$\Rightarrow 1 + \frac{(\beta)}{(B)} > 1 + \frac{(A\beta)}{(AB)}.$$

$$\begin{aligned}
 \Rightarrow & \frac{N}{(B)} > \frac{(A)}{(AB)}. \\
 \Rightarrow & \frac{N}{(A)} > \frac{(B)}{(AB)}. \\
 \Rightarrow & \frac{(A) + (\alpha)}{(A)} > \frac{(AB) + (\alpha B)}{(AB)}. \\
 \Rightarrow & 1 + \frac{(\alpha)}{(A)} > 1 + \frac{(\alpha B)}{(AB)}. \\
 \Rightarrow & \frac{(\alpha)}{(A)} > \frac{(\alpha B)}{(AB)}. \\
 \Rightarrow & \frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}, \text{ as required.}
 \end{aligned}$$

### EXERCISE 11 (a)

1. (a) Explain the following : (i) Order of a class, (ii) Ultimate classes and, (iii) Fundamental set of class frequencies.

(b) What is meant by a class-frequency of (i) first order, (ii) third order ? How would you express a class frequency of first order in terms of class frequencies of third order ?

2. What is dichotomy ? Show that the continued dichotomy according to  $n$  attributes gives rise to  $3^n$  classes.

3. (a) Given that  $(AB) = 150$ ,  $(A\beta) = 230$ ,  $(\alpha B) = 260$ ,  $(\alpha\beta) = 2,340$ ; find the other frequencies and the value of  $N$ .

(b) Given the following frequencies of the positive classes, find the frequencies of the rest of the classes :

$(A) = 977$ ,  $(AB) = 453$ ,  $(ABC) = 127$ ,  $(B) = 1,185$ ,  $(AC) = 284$ ,  $N = 12,000$ ,  $(C) = 596$ , and  $(BC) = 250$ .

Ans.  $(A\beta) = 524$ ,  $(\alpha B) = 732$ ,  $(\alpha\beta) = 10,291$ ,  $(\beta\gamma) = 935$ ,  $(\beta C) = 346$ ,  $(\beta\gamma) = 10,469$ ,  $(A\gamma) = 693$ ,  $(\alpha C) = 312$ ,  $(AB\gamma) = 326$ ,  $(\alpha BC) = 123$ ,  $(\alpha B\gamma) \approx 609$ ,  $(A\beta C) = 157$ ,  $(A\beta\gamma) = 367$ ,  $(\alpha\beta C) = 189$ ,  $(\alpha\beta\gamma) = 10,192$ .

4. Given the following data, find frequencies of (i) the remaining positive classes; and (ii) the ultimate classes :

$N = 1,800$ ,  $(A) = 850$ ,  $(B) = 780$ ,  $(C) = 326$ ,  $(AB\gamma) = 200$ ,  $(A\beta C) = 94$ ,  $(\alpha BC) = 72$ , and  $(ABC) = 50$ .

5. (a) Measurements are made on a thousand husbands and a thousand wives. If the measurements of the husbands exceed the measurements of the wives in 800 cases for one measurement, in 700 cases for another and in 660 cases for both measurements, in how many cases will both measurements on the wife exceed the measurements on the husband ?

Ans. 160

(b) An unofficial political study was made about the recent changes in Indian political scene and it was found that 919 Indira Gandhi Congress supporters and 1,268 Organisation Congress supporters wanted socialistic economy, whereas 310 Indira Gandhi Congress supporters and 503 supporters of

the Organisation Congress wanted capitalistic economy in the country. Find out the total number of Indira Gandhi's and that of the Organisation's supporters, giving the number of capitalistic economy's and of the socialistic economy's votaries, out of the individuals, who were surveyed.

6. At a competitive examination at which 600 graduates appeared, boys outnumbered girls by 96. Those qualifying for interview exceeded in number those failing to qualify by 310. The number of Science graduate boys interviewed was 300 while among the Arts graduate girls there were 25 who failed to qualify for interview. Altogether there were only 135 Arts graduates and 33 among them failed to qualify. Boys who failed to qualify numbered 18.

Find (i) the number of boys who qualified for interview,

- (ii) the total number of Science graduate boys appearing, and
- (iii) the number of Science graduate girls who qualified.

**Ans.** (i) 330, (ii) 310, and (iii) 53.

7. 100 children took three examinations A, B and C; 40 passed the first, 39 passed the second and 48 passed the third, 10 passed all the three, 21 failed all three, 9 passed the first two and failed the third, 19 failed the first two and passed the third. Find how many children passed at least two examinations. Show that for the question asked certain of the given frequencies are not necessary. Which are they ?

**Ans.** 38. Only frequencies required are  $(C)$ ,  $(\alpha\beta C)$ ,  $(AB\gamma)$ .

8. In a university examination, which was indeed very tough, 50% at least failed in "Statistics", 75% at least in Topology, 82% at least in "Functional Analysis" and 96% at least in "Applied Mathematics". How many at least failed in all the four ? (Ans. 3%)

**Hint.** Use the result in Example 11.3. Page 11.6.

9. If a collection contains  $N$  items, each of which is characterized by one or more of the attributes  $A$ ,  $B$ ,  $C$  and  $D$ , show that with the usual notations

- (i)  $(ABCD) \geq (A) + (B) + (C) + (D) - 3N$ , and
- (ii)  $(ABCD) = (ABD) + (ACD) - (AD) + (AD\beta\gamma)$ ,

where  $\beta$  and  $\gamma$  represent the characteristics of the absence of  $B$  and  $C$  respectively.

10. Given  $(A) = (\alpha) = (B) = (\beta) = \frac{1}{2}N$ ; show that  $(AB) = (\alpha\beta)$ ,  $(A\beta) = (\alpha B)$ .

11. Given that  $(A) = (\alpha) = (B) = (\beta) = (C) = (\gamma) = \frac{1}{2}N$

and also that  $(ABC) = (\alpha\beta\gamma)$ , show that  $2(ABC) = (AB) + (AC) + (BC) - \frac{1}{2}N$ .

11.6. **Consistency of Data.** Any class frequencies which have been or might have been observed within one and the same population are said to be consistent if they conform with one another and do not in any way conflict. For example; the figures  $(A) = 20$ ,  $(AB) = 25$  are inconsistent as  $(AB)$  cannot be greater than  $(A)$ , if they are observed from the same population.

'Consistency' of a set of class frequencies may be defined as the property that none of them is negative, otherwise, the data for class frequencies are said to be 'inconsistent'.

Since any class frequency can be expressed as the sum of some of the ultimate class frequencies, it is necessarily non-negative if all the ultimate class frequencies are non-negative. This provides a criterion for testing the consistency of the data. In fact, we have the following theorem.

**Theorem 11.1.** "The necessary and sufficient condition for the consistency of a set of independent class frequencies is that no ultimate class frequency is negative."

**Remark.** We can test the consistency of a set of  $2^n$  algebraically independent class frequencies by calculating the ultimate class frequencies. If any one of them is negative, the given data are inconsistent.

**11.6.1. Conditions for consistency of Data.** Criteria for consistency of class frequencies are obtained by using theorem 11.1. For a single attribute  $A$  we have conditions of consistency as follows :

$$\left. \begin{array}{l} (i) \quad (A) \geq 0 \\ (ii) \quad (\alpha) \geq 0 \Rightarrow (A) \leq N \end{array} \right\} \dots(11.5)$$

For two attributes  $A$  and  $B$ , the conditions of consistency are :

$$\left. \begin{array}{l} (i) \quad (AB) \geq 0 \\ (ii) \quad (A\beta) \geq 0 \Rightarrow (AB) \leq (A) \\ (iii) \quad (\alpha B) \geq 0 \Rightarrow (AB) \leq (B) \\ (iv) \quad (\alpha\beta) \geq 0 \Rightarrow (AB) \geq (A) + (B) - N \end{array} \right\} \dots(11.6)$$

Conditions of consistency for three attributes  $A, B$  and  $C$  are

$$\left. \begin{array}{l} (i) \quad (ABC) \geq 0 \\ (ii) \quad (AB\gamma) \geq 0 \Rightarrow (ABC) \leq (AB) \\ (iii) \quad (A\beta C) \geq 0 \Rightarrow (ABC) \leq (AC) \\ (iv) \quad (\alpha BC) \geq 0 \Rightarrow (ABC) \leq (BC) \\ (v) \quad (A\beta\gamma) \geq 0 \Rightarrow (ABC) \geq (AB) + (AC) - (A) \\ (vi) \quad (\alpha B\gamma) \geq 0 \Rightarrow (ABC) \geq (AB) + (BC) - (B) \\ (vii) \quad (\alpha\beta C) \geq 0 \Rightarrow (ABC) \geq (AC) + (BC) - (C) \\ (viii) \quad (\alpha\beta\gamma) \geq 0 \Rightarrow (ABC) \leq (AB) + (BC) + (AC) - (A) - (B) - (C) + N \end{array} \right\} \dots(11.7)$$

(i) and (viii) in (11.7) give :

$$(AB) + (BC) + (AC) \geq (A) + (B) + (C) - (N)$$

Similarly

$$\left. \begin{array}{l} (ii) \text{ and } (vii) \Rightarrow (AC) + (BC) - (AB) \leq (C) \\ (iii) \text{ and } (vi) \Rightarrow (AB) + (BC) - (AC) \leq (B) \\ (iv) \text{ and } (v) \Rightarrow (AB) + (AC) - (BC) \leq (A) \end{array} \right\} \dots(11.8)$$

**Remark.** As already pointed out [c.f. Remarks (3) and (4), § 11.4.2)],  $2^n$  algebraically independent class frequencies are necessary to specify the data completely, one such set being the set of ultimate class frequencies and the other being the set of positive class frequencies. If the data supplied are incomplete so that it is not possible to determine all the class frequencies, then the conditions (11.5), (11.6) and (11.8) for one, two and three attributes respectively, enable us to assign the limits within which an unknown class frequency can lie.

**Example 11.5.** Examine the consistency of the following data :

$N = 1,000$ ,  $(A) = 600$ ,  $(B) = 500$ ,  $(AB) = 50$ , the symbols having their usual meaning.

**Solution.** We have

$$(A\beta) = N - (A) - (B) + (AB) = 1000 - 600 - 500 + 50 = -50.$$

Since  $(A\beta) < 0$ , the data are inconsistent.

**Example 11.6.** Among the adult population of a certain town 50 per cent are males, 60 per cent are wage earners and 50 per cent are 45 years of age or over, 10 per cent of the males are not wage-earners and 40 per cent of the males are under 45. Make the best possible inference about the limits within which the percentage of persons (male or female) of 45 years or over are wage-earners.

**Solution.** Let  $N = 100$ . Then denoting males by  $A$ , wage-earners by  $B$  and 45 years of age or over by  $C$ , we are given :

$$N = 100, (A) = 50, (B) = 60, (C) = 50$$

$$(A\beta) = \frac{10}{100} \times 50 = 5, (A\gamma) = \frac{40}{100} \times 50 = 20$$

$$\therefore (AB) = (A) - (A\beta) = 45, (AC) = (A) - (A\gamma) = 30$$

We are required to find the limits for  $(BC)$ .

Conditions of consistency (11.8) give

$$(i) \quad (AB) + (BC) + (AC) \geq (A) + (B) + (C) - N$$

$$\Rightarrow (BC) \geq 50 + 60 + 50 - 100 - 45 - 30 = -15$$

$$(ii) (AB) + (AC) - (BC) \leq A$$

$$\Rightarrow (BC) \geq (AB) + (AC) - (A) = 45 + 30 - 50 = 25$$

$$(iii) (AB) + (BC) - (AC) \leq (B)$$

$$\Rightarrow (BC) \leq (B) + (AC) - (AB) = 60 + 30 - 45 = 45$$

$$(iv) (AC) + (BC) - (AB) \leq (C)$$

$$(BC) \leq (C) + (AB) - (AC) = 50 + 45 - 30 = 65$$

$$(i) \text{ to } (iv) \Rightarrow 25 \leq (BC) \leq 45$$

Hence the percentage of wage-earning population of 45 years or over must lie between 25 and 45.

**Example 11.7.** In a series of houses actually invaded by smallpox, 70% of the inhabitants are attacked and 85% have been vaccinated. What is the lowest percentage of the vaccinated that must have been attacked?

**Solution.** Let  $A$  and  $B$  denote the attributes of the inhabitants being attacked and vaccinated respectively. Then we are given :

$$N = 100, (A) = 70 \text{ and } (B) = 85$$

Consistency condition gives :

$$(AB) \geq (A) + (B) - N \Rightarrow (AB) \geq 55$$

Hence the lowest percentage of inhabitants vaccinated, who have been attacked is

$$\frac{(AB)}{(B)} \times 100 = \frac{55}{85} \times 100 = 64.7\%$$

**Example 11.8.** Show that if

$$\frac{(A)}{N} = x, \frac{(B)}{N} = 2x, \frac{(C)}{N} = 3x$$

and  $\frac{(AB)}{N} = \frac{(BC)}{N} = \frac{(CA)}{N} = y,$

then the value of neither  $x$  nor  $y$  can exceed 1/4.

**Solution.** Conditions of consistency give :

$$(AB) \leq (A) \Rightarrow Ny \leq Nx \Rightarrow y \leq x \quad \dots(i)$$

Also  $(BC) \geq (B) + (C) - N$

$$\Rightarrow \frac{(BC)}{N} \geq \frac{(B)}{N} + \frac{(C)}{N} - 1$$

$$\Rightarrow y \geq 2x + 3x - 1$$

$$\Rightarrow 5x - 1 \leq y \quad \dots(ii)$$

(i) and (ii) give

$$5x - 1 \leq x \Rightarrow 4x \leq 1 \Rightarrow x \leq \frac{1}{4} \quad \dots(iii)$$

Thus from (i) and (iii) we have  $y \leq x \leq \frac{1}{4}$ , which establishes the result.

**Example 11.9.** Show that (i) If all A's are B's and all B's are C's then all A's are C's, (ii) If all A's are B's and no B's are C's then no A's are C's.

**Solution.** (i) All A's are B's  $\Rightarrow (AB) = (A)$   
and all B's are C's  $\Rightarrow (BC) = (B)$  }  $\dots(*)$

To prove  $(AC) = (A)$

We have  $(AB) + (BC) - (AC) \leq (B)$

$$\Rightarrow (A) + (B) - (AC) \leq (B) \quad [\text{Using } (*)]$$

$$\Rightarrow (A) \leq (AC) \Rightarrow (AC) \geq (A)$$

But since  $(AC) \neq (A)$ , we have  $(AC) = (A)$ , as desired.

(ii) We are given  $(AB) = (A)$  and  $(BC) = 0$  and we want to prove  $(AC) = 0$ .

We have

$$(AB) + (AC) - (BC) \leq (A)$$

$$\Rightarrow (A) + (AC) - 0 \leq (A)$$

$$\Rightarrow (AC) \leq 0$$

And since  $(AC) \geq 0$ , we must have  $(AC) = 0$ .

### EXERCISE 11(b)

1. What do you understand by consistency of given data ? How do you check it ?

2. (a) If a report gives the following frequencies as actually observed, show that there must be a misprint or mistake of some sort, and that possibly the misprint consists in the dropping of 1 before 85 given as the frequency  $(BC)$ :

$$N = 1000, (A) = 510, (B) = 490, (C) = 427, (AB) = 189, (AC) = 140, (BC) = 85.$$

(b) A student reported the results of a survey in the following manner, in terms of the usual notations :

$N = 1000$ ,  $(A) = 525$ ,  $(B) = 312$ ,  $(C) = 470$ ,  $(AB) = 42$ ,  $(BC) = 86$ ,  $(AC) = 147$ , and  $(ABC) = 25$ .

Examine the consistency of the above data.

(c) Examine the consistency and adequacy of the following data to determine the frequencies of the remaining positive and ultimate classes.

$N = 10,000$ ,  $(A) = 1087$ ,  $(B) = 286$ ,  $(C) = 877$ ,

$(CA\beta) = 281$ ,  $(C\alpha\beta) = 86$ ,  $(\gamma AB) = 78$ ,  $(ABC) = 57$

3. Given that  $(A) = (B) = (C) = \frac{1}{2}N$  and 80 per cent of  $A$ 's are  $B$ 's, 75 per cent of  $A$ 's are  $C$ 's, find the limits to the percentage of  $B$ 's that are  $C$ 's.

Ans. 55% and 95%.

4. If  $(A) = 50$ ,  $(B) = 60$ ,  $(C) = 50$ ,  $(A\beta) = 5$ ,  $(A\gamma) = 20$ ,  $N = 100$ , find the greatest and the least possible values of  $(BC)$  so that the data may be consistent.

Ans.  $25 \leq (BC) \leq 45$

5. If  $1,000 = N = 1\frac{5}{3}(A) = 2(B) = 2\frac{5}{2}(C) = 5(AB)$ , and  $(AC) = (BC)$ , what should be the minimum value of  $(BC)$  ?

Ans. 150

6. Given that  $(A) = (B) = (C) = \frac{1}{2}N = 50$  and  $(AB) = 30$ ,  $(AC) = 25$ , find the limits within which  $(BC)$  will lie.

7. In a university examination 65% of the candidate passed in English, 90% passed in the second language and 60% passed in the optional subjects. Find how many at least should have passed the whole examination.

Ans: 15%. Hint. Use Example 11-3.

8. A market investigator returns the following data. Of 1,000 people consulted 811 liked chocolates, 752 liked toffees and 418 liked boiled sweets, 570 liked both chocolates and toffees, 356 liked chocolates and boiled sweets and 348 liked toffees and boiled sweets, 297 liked all three. Show that this information as it stands must be incorrect.

9. (a) In a school, 50 per cent of the students are boys, 60 per cent are Hindus and 50 per cent are 10 years of age or over. Twenty per cent of the boys are not Hindus and 40 per cent of the boys are under 10. What conclusions can you draw in regard to percentage of Hindu students of 10 years or over ?

(b) In a college, 50 per cent of the students are boys, 60 per cent of the student are above 18 years and 80 per cent receive scholarships. 35 per cent of the students are boys above 18 years of age, 45 per cent are boys receiving scholarships, and 42 per cent are above 18 years and receive scholarships. Determine the limits to the proportion of boys above 18 years who are in receipt of scholarships.

Ans. Between 30 and 32.

10. The following summary appears in a report on a survey covering 1,000 fields. Scrutinise the numbers and point out if there is any mistake or misprint in them.

Manured fields	510
Irrigated fields	490
Fields growing improved varieties	427
Fields both irrigated and manured	189
Fields both manured and growing improved varieties	140
Fields both irrigated and growing improved varieties	85

Hint. Let  $A$  : manured fields;

$B$  : Irrigated fields

and  $C$  : Growing improved varieties; then  $(\alpha\beta\gamma) < 0$ .

11. A social survey in a village revealed that there were more uneducated employed males than educated ones; there were more educated employed males than uneducated unemployed males. There were more educated unemployed under 35 years of age than employed uneducated males over 35 years of age. Show that there are more uneducated employed males under 35 years of age than educated unemployed males over 35 years of age.

12. In a war between White and Red forces, there are more Red soldiers than White, there are more armed Whites than unarmed Reds, there are fewer armed Reds with ammunition than unarmed Whites without ammunition. Show that there are more armed Reds without ammunition than unarmed Whites with ammunition.

13. Given that  $(A) = (B) = (C) = \frac{N}{2}$ ,  $\frac{(AB)}{N} = \frac{(AC)}{N} = p$ , find what must be the greatest and least values of  $p$  in order that we may infer that  $(BC)/N$ , exceeds any given value, say  $q$ .

$$\text{Ans. } \frac{1}{4}(1 - 2q) \leq p \leq \frac{1}{4}(1 + 2q).$$

11.7. Independence of Attributes. Two attributes  $A$  and  $B$  are said to be independent if there exists no relationship of any kind between them. If  $A$  and  $B$  are independent, we would expect (i) the same proportion of  $A$ 's amongst  $B$ 's as amongst  $\beta$ 's, (ii) the proportion of  $B$ 's amongst  $A$ 's is same as that amongst the  $\alpha$ 's. For example, if insanity and deafness are independent, the proportion of the insane people among deafs and non-deafs must be same.

11.7.1. Criterion of Independence. If  $A$  and  $B$  are independent, then (i) in § 11.7 gives

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \quad \dots(11.9)$$

$$\Rightarrow 1 - \frac{(AB)}{(B)} = 1 - \frac{(A\beta)}{(\beta)}$$

$$\Rightarrow \frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)} \quad \dots(11.9a)$$

Similarly, (ii) in § 11.7 gives

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)} \quad \dots(11.10)$$

$$\Rightarrow 1 - \frac{(AB)}{(A)} = 1 - \frac{(\alpha B)}{(\alpha)}$$

$$\Rightarrow \frac{(AB)}{(A)} = \frac{(\alpha\beta)}{(\alpha)} \quad \dots(11.10a)$$

In fact (11.9)  $\Rightarrow$  (11.10) and vice-versa.

For example, (11.9) gives

$$\begin{aligned} \frac{(AB)}{(B)} &= \frac{(A\beta)}{(\beta)} = \frac{(AB) + (A\beta)}{(B) + (\beta)} = \frac{(A)}{(N)} \\ \Rightarrow \frac{(AB)}{(A)} &= \frac{(B)}{N} = \frac{(B) - (AB)}{N - (A)} = \frac{(\alpha B)}{(\alpha)}, \end{aligned} \quad \dots(11.10b)$$

which is (11.10). Similarly, starting from (11.10), we would arrive at (11.9).

It becomes easier to grasp the nature of the above relations if the frequencies are supposed to be grouped into a table with two rows and two columns as follows.:

Attributes	$A$	$\alpha$	Total
$B$	$(AB)$	$(\alpha B)$	$(B)$
$\beta$	$(A\beta)$	$(\alpha\beta)$	$(\beta)$
Total	$(A)$	$(\alpha)$	$N$

Second criterion of independence may be obtained in terms of the class frequencies of first order. (11.10b) gives

$$\frac{(AB)}{(A)} = \frac{(A)(B)}{N} \quad \dots(11.11)$$

$$\Rightarrow \frac{(AB)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N} \quad \dots(11.11a)$$

which leads to the following important fundamental rule :

*"If the attributes A and B are independent, the proportion of AB's in the population is equal to the product of the proportions of A's and B's in the population."*

We may obtain a third criterion of independence in terms of second order class frequencies, as follows.

$$(AB) \cdot (\alpha\beta) = \frac{(A)(B)}{N} \cdot \frac{(\alpha)(\beta)}{N} = \frac{(A)(\beta)}{N} \cdot \frac{(\alpha)(B)}{N} \quad (\text{Using } 11.11)$$

$$\Rightarrow \left. \begin{aligned} (AB) \cdot (\alpha\beta) &= (A\beta) \cdot (\alpha B) \\ \frac{(AB)}{(\alpha B)} &= \frac{(A\beta)}{(\alpha\beta)} \end{aligned} \right\} \quad \dots(11.12)$$

Aliter. (11.12) may also be obtained from (11.9) and (11.9a) as explained below:

$$(11.9) \text{ and } (11.9a) \Rightarrow \frac{(AB)}{(A\beta)} = \frac{(B)}{(\beta)} = \frac{(\alpha B)}{(\alpha\beta)}$$

$$\Rightarrow (AB)(\alpha\beta) = (A\beta) \cdot (\alpha\beta)$$

Similarly, (11.10) and (11.10a) give the same result.

### 11.7.2. Symbols $(AB)_0$ and $\delta$ . Let us write

$$(AB)_0 = \frac{(A)(B)}{N} \quad \dots(11.13)$$

which is the value of  $(AB)$  under the hypothesis that the attributes  $A$  and  $B$  are independent.

$$\text{Let } \delta = (AB) - (AB)_0 \quad \dots(11.14)$$

denote the excess of  $(AB)$  over  $(AB)_0$ . Then

$$\begin{aligned}\delta &= (AB) - \frac{(A)(B)}{N} = \frac{1}{N} [N(AB) - (A)(B)] \\ &= \frac{1}{N} \left[ \{(AB) + (A\beta) + (\alpha B) + (\alpha\beta)\} (AB) \right. \\ &\quad \left. - \{(AB) + (A\dot{\beta})\} \{(AB) + (\alpha B)\} \right] \\ &= \frac{1}{N} \left[ (AB)(\alpha\beta) - (A\beta)(\alpha B) \right] \quad [\text{On simplification}]\end{aligned}$$

$$(11.12) \Rightarrow \delta = 0, \text{ if } A \text{ and } B \text{ are independent.} \quad \dots(11.15)$$

**Example 11.10.** If  $\delta = (AB) - (AB)_0$ , then with usual notations, prove that

$$(i) \quad [(A - (\alpha))[(B) - (\beta)] + 2N\delta = (AB)^2 + (\alpha\beta)^2 - (A\beta)^2 - (\alpha B)^2$$

$$(ii) \quad \delta = \frac{(B)(\beta)}{N} \left\{ \frac{(AB)}{(B)} - \frac{(A\beta)}{(\beta)} \right\} = \frac{(A)(\alpha)}{N} \left\{ \frac{(AB)}{(A)} - \frac{(\alpha B)}{(\alpha)} \right\}$$

**Solution.** (i) We have  $\delta = (AB) - (AB)_0 = (AB) - \frac{(A)(B)}{N}$

$$(ii) \frac{(B)(\beta)}{N} \left[ \frac{(AB)}{\beta} - \frac{(A\beta)}{B} \right] = \frac{1}{N} [(\beta \cdot (AB)) - (B)(A\beta)]$$

$$\begin{aligned}
 &= \frac{1}{N} \left[ (AB) \{ N - (B) \} - (B) \{ (A) - (AB) \} \right] \\
 &= \frac{1}{N} \left[ N \cdot (AB) - (A) \cdot (B) \right] = (AB) - \frac{(A) \cdot (B)}{N} = \delta
 \end{aligned}$$

Since  $\delta$  is symmetric in  $A$  and  $B$ , by interchanging  $A$  and  $B$ , we will obtain the second result.

**11-8. Association of Attributes.** Two attributes  $A$  and  $B$  are said to be associated if they are not independent but are related in some way or the other. They are said to be

$$\left. \begin{array}{l} \text{positively associated if } (AB) > \frac{(A)(B)}{N} \\ \text{and} \quad \text{negatively associated if } (AB) < \frac{(A)(B)}{N} \end{array} \right\} \dots(11-16)$$

In other words, two attributes  $A$  and  $B$  are positively associated if  $\delta > 0$ , negatively associated if  $\delta < 0$  or and are independent if  $\delta = 0$  (c.f. § 11-7-2).

**Remarks 1.** Two attributes  $A$  and  $B$  are said to be *completely associated* if  $A$  cannot occur without  $B$ , though  $B$  may occur without  $A$  and *vice-versa*. In other words, for complete association either all  $A$ 's are  $B$ 's i.e.,  $(AB) = (A)$  or all  $B$ 's are  $A$ 's i.e.,  $(AB) = (B)$  according as either  $A$ 's or  $B$ 's are in a minority. Similarly, *complete dissociation* means that no  $A$ 's are  $B$ 's i.e.,  $(AB) = 0$  or no  $\alpha$ 's are  $\beta$ 's i.e.,  $(\alpha\beta) = 0$  or more generally when either of these statements is true.

2. It should be carefully noted that the word 'association' used in Statistics is technically different from the general notion of association as used in day-to-day life. Ordinarily, two attributes are said to be associated if they occur together in a number of cases. But statistically two attributes are said to be associated if they occur together in a large number of cases than expected if they were independent, i.e., if  $\delta = (AB) - (A)(B)/N > 0$ . In Statistics, the statement that "some  $A$ 's are  $B$ 's", however great the proportion, does not necessarily imply association between them. Thus to find out if two attributes are associated, we must know  $(A)$ ,  $(B)$ ,  $(AB)$  and  $N$ . Incomplete information will not enable us to conclude anything about association between them. For example, consider the following statement :

"90 per cent of the people who drink alcohol die before reaching the age of 75 years. Hence drinking is bad for longevity of life."

The inference drawn is not correct, since the given information is not complete for drawing any valid conclusions about association. It might happen that 95% of the people who do not drink, die before reaching 75 years of age. In that case drinking might be found good for longevity of life.

3. *Sampling fluctuations.* If  $\delta \neq 0$  and its value is fairly small, then it is possible that this association is just by chance (or commonly termed as 'due to fluctuations of sampling') and not really significant of any real association between the attributes. We should not, therefore, draw hasty conclusions about association or dissociation unless  $\delta$ , the difference between  $(AB)$  and its expected value (under the hypothesis of independence)  $(A)(B)/N$ , is significant. The

problem : 'how much difference is to be regarded as significant' will be discussed in detail in Chapters 12 (Large sample test for attributes) and 13 (Chi-square test of goodness of fit). This point has been raised here only as a precautionary measure to warn the reader against drawing hasty inferences.

**11.8.1. Yule's Coefficient of Association.** As a measure of the intensity of association between two attributes  $A$  and  $B$ , G. Udny Yule gave the coefficient of association  $Q$ , defined as follows :

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \quad \dots(11.17)$$

If  $A$  and  $B$  are independent,  $\delta = 0 \Rightarrow Q = 0$ .

If  $A$  and  $B$  are completely associated, then

either  $(AB) = (A) \Rightarrow (A\beta) = 0$   
 or  $(AB) = (B) \Rightarrow (\alpha B) = 0$

and in each case  $Q = +1$ .

If  $A$  and  $B$  are in complete dissociation then either  $(AB) = 0$  or  $(\alpha\beta) = 0$  and we get  $Q = -1$ .

Hence  $-1 \leq Q \leq 1 \quad \dots(11.18)$

**Remark.** An important property of  $Q$  is that it is independent of the relative proportion of  $A$ 's or  $\alpha$ 's in the data. Thus if all the terms containing  $A$  in  $Q$  are multiplied by a constant,  $k$  (say), its value remains unaltered. Similarly for  $B$ ,  $\beta$  and  $\alpha$ . This property renders it specially useful to situations where the proportions are arbitrary, e.g., experiments.

**11.8.2. Coefficient of Colligation.** Another coefficient with the same properties as  $Q$ , is the coefficient of colligation  $Y$ , given by

$$Y = \left\{ 1 - \sqrt{\frac{(AB)(\alpha B)}{(AB)(\alpha\beta)}} \right\} \Big/ \left\{ 1 + \sqrt{\frac{(AB)(\alpha B)}{(AB)(\alpha\beta)}} \right\} \quad \dots(11.19)$$

**Remarks.** 1. Obviously  $Q = 0 \Rightarrow Y = \frac{1 - 1}{1 + 1} = 0$ .

$Q = -1 \Rightarrow Y = -1$  and  $Q = 1 \Rightarrow Y = 1$  and conversely.

2. If we let  $\frac{(AB)(\alpha B)}{(AB)(\alpha\beta)} = k$ , so that

$$\begin{aligned} Y &= \frac{1 - \sqrt{k}}{1 + \sqrt{k}} \Rightarrow Y^2 = \frac{1 + k - 2\sqrt{k}}{1 + k + 2\sqrt{k}} \\ \Rightarrow 1 + Y^2 &= \frac{2(1 + k)}{1 + k + 2\sqrt{k}} = \frac{2(1 + k)}{(1 + \sqrt{k})^2} \\ \therefore \frac{2Y}{1 + Y^2} &= \frac{2(1 - \sqrt{k})(1 + \sqrt{k})}{2(1 + k)} = \frac{1 - k}{1 + k} \\ &= \frac{1 - \frac{(AB)(\alpha B)}{(AB)(\alpha\beta)}}{1 + \frac{(AB)(\alpha B)}{(AB)(\alpha\beta)}} = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \end{aligned}$$

$$\Rightarrow Q = \frac{2Y}{1 + Y^2} \quad \dots(11.20)$$

**Example 11.11.** Find if A and B are independent, positively associated or negatively associated, in each of the following cases :

(i)  $N = 1000$ ,  $(A) = 470$ ,  $(B) = 620$ , and  $(AB) = 320$ .

(ii)  $(A) = 490$ ,  $(AB) = 294$ ,  $(\alpha) = 570$ , and  $(\alpha B) = 380$ .

(iii)  $(AB) = 256$ ,  $(\alpha B) = 768$ ,  $(A\beta) = 48$ , and  $(\alpha\beta) = 144$ .

**Solution.** (i)  $\delta = (AB) - \frac{(A)(B)}{N}$

$$= 320 - \frac{470 \times 620}{1000} = 320 - 291.4 = 28.6$$

Since  $\delta > 0$ , A and B are positively associated.

(ii) We have  $N = (A) + (\alpha) = 490 + 570 = 1060$

$$(B) = (AB) + (\alpha B) = 294 + 380 = 674$$

$$\therefore \delta = (AB) - \frac{(A)(B)}{N} = 294 - \frac{490 \times 674}{1060} = 294 - 311.6 < 0$$

Hence A and B are negatively associated.

(iii)  $(A) = (AB) + (A\beta) = 256 + 48 = 304$

$$(B) = (AB) + (\alpha B) = 256 + 768 = 1024$$

$$N = (AB) + (A\beta) + (\alpha B) + (\alpha\beta) = 256 + 48 + 768 + 144 = 1216$$

$$\therefore \delta = (AB) - \frac{(A)(B)}{N} = 256 - \frac{304 \times 1024}{1216} = 0$$

Hence A and B are independent.

**Aliter.** Since all the four frequencies of order 2 are given, using (11.15), we have

$$\begin{aligned} \delta &= \frac{1}{N} [(AB)(\alpha\beta) - (A\beta)(\alpha B)] = \frac{1}{N} [256 \times 144 - 48 \times 768] \\ &= \frac{256}{N} [144 - 48 \times 3] = 0 \end{aligned}$$

$\Rightarrow$  A and B are independent.

**Example 11.12.** Investigate the association between darkness of eye-colour in father and son from the following data :

Fathers with dark eyes and sons with dark eyes : 50

Fathers with dark eyes and sons with not dark eyes : 79

Fathers with not dark eyes and sons with dark eyes : 89

Fathers with not dark eyes and sons with not dark eyes : 782

Also tabulate for comparison the frequencies that would have been observed had there been no heredity.

**Solution.** Let A : Dark eye-colour of father and

B : Dark eye-colour of son.

Then we are given  $(AB) = 50$ ,  $(A\beta) = 79$ ,  $(\alpha B) = 89$ ,  $(\alpha\beta) = 782$

$$\therefore Q = \frac{50 \times 782 - 79 \times 89}{50 \times 782 + 79 \times 89} = \frac{32069}{46131} = +0.69$$

Hence there is a fairly high degree of positive association between the eye colour of fathers and sons.

$$\text{We have } (A) = (AB) + (A\beta) = 50 + 79 = 129$$

$$(B) = (AB) + (\alpha B) = 50 + 89 = 139$$

$$(\alpha) = (\alpha B) + (\alpha\beta) = 89 + 782 = 871$$

$$(\beta) = (A\beta) + (\alpha\beta) = 79 + 782 = 861$$

$$N = (A) + (\alpha) = 129 + 871 = 1000$$

Under the condition of *no heredity*, i.e., independence of attributes  $A$  and  $B$ , we have

$$(AB)_0 = \frac{(A)(B)}{N} = \frac{129 \times 139}{1000} = 18; (A\beta)_0 = \frac{(A)(\beta)}{N} = \frac{129 \times 861}{1000} = 111$$

$$(\alpha B)_0 = \frac{(\alpha)(B)}{N} = \frac{871 \times 139}{1000} = 121; (\alpha\beta)_0 = \frac{(\alpha)(\beta)}{N} = \frac{871 \times 861}{1000} = 750$$

**Example 11.13.** Can vaccination be regarded as a preventive measure for small pox from the data given below?

'Of 1482 persons in a locality exposed to small-pox, 368 in all were attacked.'

'Of 1482 persons, 343 had been vaccinated and of these only 35 were attacked.'

**Solution.** Let  $A$  denote the attribute of vaccination and  $B$  that of attack by small-pox. Then the given data are :

$$N = 1482, (A) = 368, (B) = 343 \text{ and } (AB) = 35$$

$$(\alpha\beta) = N - (A) - (B) + (AB) = 1482 - 368 - 343 + 35 = 806$$

$$(A\beta) = (A) - (AB) = 368 - 35 = 333$$

$$(\alpha B) = (B) - (AB) = 343 - 35 = 308$$

$$\therefore Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (a\beta)(\alpha B)} = \frac{35 \times 806 - 333 \times 308}{35 \times 806 + 333 \times 308} = -0.57$$

Thus, there is negative association between  $A$  and  $B$  i.e., between 'attacked' and 'vaccinated'. In other words, there is positive association between not attacked and vaccinated. Hence vaccination can be regarded as a preventive measure for smallpox.

### EXERCISE 11(c)

1. (a) What do you mean by independence of attributes? Give a criterion of independence for attributes  $A$  and  $B$ .

(b) What are the various methods of finding whether two attributes are associated, dissociated or independent? Deduce any one such measure of association.

(c) When are two attributes said to be positively associated and negatively associated? Also define complete association and dissociation of two attributes.

(d) Derive an expression for a measure of association between two attributes.

(e) What is association of attributes ? Write a note on the strength of association and how it is measured ?

(f) Find whether the attributes  $\alpha$  and  $\beta$  are positively associated, negatively associated or independent. Given  $(AB) = 500$ ,  $(\alpha) = 800$ ,  $(B) = 600$ ,  $N = 1500$ .

2. (a) Define Yule's coefficient of association and the coefficient of Colligation. Establish the following relation between coefficient of association  $Q$  and coefficient of colligation  $Y$  :

$$Q = \frac{2Y}{1 + Y^2}$$

(b) For the following table, give Yule's coefficient of association ( $Q$ ) and coefficient of Colligation ( $Y$ ). Examine the cases (i)  $bc = 0$ , (ii)  $ad = 0$ , and (iii)  $ad = bc$ .

	<i>B</i>	<i>not B</i>
<i>A</i>	<i>a</i>	<i>b</i>
<i>not A</i>	<i>c</i>	<i>d</i>

Ans.  $Q = 1 = Y$  if  $bc = 0$  and  $Q = -1 = Y$  if  $ad = 0$  and  $Q = 0$  if  $ad = bc$ .

(c) Prove that in the usual notations  $Q = 2Y/(1 + Y^2)$ . What is the range of values for  $Q$  ?

(d) If an attribute  $A$  is known to be completely associated with an attribute  $B$ , (i) what can you infer about the association between  $\alpha$  and  $\beta$  ? ( $\alpha$  and  $\beta$  are equivalent to 'not  $A$ ' and 'not  $B$ ' respectively), (ii)  $\alpha$  and  $B$  ?

3. (a) The following table is reproduced from a memoir written by Karl Pearson :

<i>Eye colour in son</i>			
	<i>Not light</i>	<i>Light</i>	
<i>Eye colour in father</i>	<i>Not light</i>	230	148
	<i>Light</i>	151	471

Discuss if the colour of son's eyes is associated with that of father.

Ans. Yes. Positively associated,  $Q = 0.66$ .

(b) The following table shows the result of inoculation against cholera.

	<i>Not attacked</i>	<i>Attacked</i>
<i>Inoculated</i>	431	5
<i>Not-inoculated</i>	291	9

Examine the effect of inoculation in controlling susceptibility to cholera.

Ans. Inoculation is effective in controlling cholera.

4. (a) find the association between proficiency in English and in Hindi among candidates at a certain test if 245 of them passed in Hindi, 285 failed in Hindi, 190 failed in Hindi but passed in English and 147 passed in both.

(b) The male population of a state is 250 lakhs. The number of literate males is 20 lakhs and total number of male criminals is 26 thousand. The number of literate male criminals is 2 thousand. Do you find any association between literacy and criminality ?

Ans. Literacy and criminality are positively associated.

(c) From the following particulars find whether blindness and baldness are associated :

Total population	1,62,64,000
Number of baldheaded	24,441
Number of blind	7,263
Number of baldheaded blind	221

5. In a certain investigation carried on with regard to 500 graduates and 1500 non-graduates, it was found that the number of employed graduates was 450 while the number of unemployed non-graduates was 300. In the second investigation 5000 cases were examined. The number of non-graduates was 3000 and the number of employed non-graduates was 2500. The number of graduates who were found to be employed was 1600:

Calculate the coefficient of association between graduation and employment in both the investigations.

Can any definite conclusion be drawn from the coefficients ?

Ans.  $Q$  (1st Investigation) = +0.38,  $Q$  (Second Investigation) = -0.11

6. (a) Three aptitude tests  $A$ ,  $B$ ,  $C$  were given to 200 apprentice trainees. From amongst them 80 passed test  $A$ , 78 passed test  $B$  and 96 passed the third test. While 20 passed all the three tests, 42 failed all the three, 18 passed  $A$  and  $B$  but failed  $C$  and 38 failed  $A$  and  $B$  but passed the third. Determine (i) how many trainees passed at least two of the three tests and (ii) whether the performances in tests  $A$  and  $B$  are associated. Ans. (i) 76, (ii)  $Q = 0.3$

(b) In a survey of a population of 12000, information is gathered regarding three attributes  $A$ ,  $B$  and  $C$ . In the usual notations,

$$(A) = 980, (AB) = 450, (ABC) = 130$$

$$(B) = 1190, (AC) = 280, (C) = 600 \text{ and } (BC) = 250.$$

Find : (i)  $(\alpha\beta\gamma)$  (ii)  $Q_{AB}$  = Coefficient of Association between  $A$  and  $B$ .

Comment on your findings.

7. A group of 1000 fathers was studied and it was found that 12.9% had dark eyes. Among them the ratio of those having sons with dark eyes to those having sons with not dark eyes was 1 : 1.58. The number of cases where fathers and sons both did not have dark eyes was 782. Calculate coefficient of association between darkness of eye colour in father and son. Give the frequencies that would have been observed had there been completely no heredity.

Hint.  $(AB) = 50$ ,  $(A\bar{B}) = 79$ ,  $(\bar{A}B) = 89$  and  $(\bar{A}\bar{B}) = 782$ .

8. A census revealed the following figures of the blind and the insane in two age-groups in a certain population :

	Age-Group 15—25 years	Age-Group over 25 years
Total population	2,70,000	1,60,200
Number of blind	1,000	2,000
Number of insane	6,000	1,000
Number of insane among the blind	19	9

(i) Obtain a measure of association between blindness and insanity for each age-group.

(ii) Which group shows more association or dis-association (if any) ?

9. Show that if  $(AB)_1, (\alpha B)_1, (A\beta)_1, (\alpha\beta)_1$  and  $(AB)_2, (\alpha B)_2, (A\beta)_2$  and  $(\alpha\beta)_2$  be two aggregates corresponding to the same values of  $(A), (B), (\alpha)$  and  $(\beta)$ , then

$$(AB)_1 - (AB)_2 = (\alpha B)_2 - (\alpha B)_1 = (A\beta)_2 - (A\beta)_1 = (\alpha\beta)_1 - (\alpha\beta)_2$$

10. Show that if  $\delta = (AB) - \frac{(A)(B)}{N}$ , then

$$\delta = \frac{1}{N} [(AB)(\alpha\beta) - (A\beta)(\alpha B)]$$

### OBJECTIVE TYPE QUESTIONS

1. State, giving reasons, whether each of the following statements is true or false :

(i) There is no difference between correlation and association.

(ii) All the class frequencies of various orders are independent of each other.

(iii) If the attributes  $A$  and  $B$  are positively associated, then  $\alpha$  and  $B$  are also positively associated.

(iv) Square of Yule's coefficient of association cannot exceed 1.

(v) Yule's coefficient of association cannot be negative.

(vi) For two attributes  $A$  and  $B$ , the coefficient of association  $Q$  is 0.36. If each ultimate class frequency is doubled then  $Q$  is 0.72.

(vii) If  $(AB) = 10, (\alpha B) = 15, (A\beta) = 20$  and  $(\alpha\beta) = 30$ , then  $A$  and  $B$  are associated.

(viii) If every item which possesses an attribute  $A$  possesses the attribute  $B$  as well, then the coefficient of association between  $A$  and  $B$  is 1.

II. Indicate the correct answer :

(i) In case of two attributes  $A$  and  $B$ , the ultimate class frequencies are :-

(a) :  $(A), (b) : (AB), (c) : (\alpha), (d) : (B)$ .

(ii) The condition for the consistency of a set of independent class frequencies is that no ultimate class frequency is (a) zero, (b) positive, (c) negative.

(iii) Attributes  $A$  and  $B$  are said to be independent if

(a)  $(AB) > \frac{(A) \times (B)}{N}$ , (b)  $(AB) = \frac{(A) \times (B)}{N}$ , (c)  $(AB) < \frac{(A) \times (B)}{N}$

(iv) Attributes  $A$  and  $B$  are said to be positively associated if

(a)  $\frac{(AB)}{(B)} < \frac{(A\beta)}{(\beta)}$ , (b)  $\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$ , (c)  $\frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)}$ , (d)  $\frac{(AB)}{(A)} < \frac{(A\beta)}{(\beta)}$

(v) If  $N = 50, (A) = 35, (B) = 25, (AB) = 15$ , then the attributes  $A$  and  $B$  are said to be :

(a) correlated, (b) independent, (c) negatively associated, (d) positively associated

(vi) When there is a perfect positive association between two attributes,  $Q$  would be (a) zero, (b) - 0.9, (c) -1, (d) +1.

## ***Sampling and Large Sample Tests***

---

**12·1. Sampling—Introduction.** Before giving the notion of sampling we will first define *population*. In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. This group of individuals under study is called *population or universe*. Thus in statistics, population is an aggregate of objects, animate or inanimate, under study. The population may be finite or infinite.

It is obvious that for any statistical investigation complete enumeration of the population is rather impracticable. For example, if we want to have an idea of the average per capita (monthly) income of the people in India, we will have to enumerate all the earning individuals in the country, which is rather a very difficult task.

If the population is infinite, complete enumeration is not possible. Also if the units are destroyed in the course of inspection (e.g., inspection of crackers, explosive materials, etc.), 100% inspection, though possible, is not at all desirable. But even if the population is finite or the inspection is not destructive, 100% inspection is not taken recourse to because of multiplicity of causes, viz., administrative and financial implications, time factor, etc., and we take the help of *sampling*.

A finite subset of statistical individuals in a population is called a *sample* and the number of individuals in a sample is called the *sample size*.

For the purpose of determining population characteristics, instead of enumerating the entire population, the individuals in the sample only are observed. Then the sample characteristics are utilised to approximately determine or estimate the population. For example, on examining the sample of a particular stuff we arrive at a decision of purchasing or rejecting that stuff. The error involved in such approximation is known as *sampling error* and is inherent and unavoidable in any and every sampling scheme. But sampling results in considerable gains, especially in time and cost not only in respect of making observations of characteristics but also in the subsequent handling of the data.

Sampling is quite often used in our day-to-day practical life. For example, in a shop we assess the quality of sugar, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

**12·2. Types of Sampling.** Some of the commonly known and frequently used types of sampling are :

- (i) *Purposive sampling.* (ii) *Random sampling.* (iii) *Stratified sampling.*  
 (iv) *Systematic Sampling.*

Below we will precisely explain these terms, without entering into detailed discussion.

**12-2-1. Purposive Sampling.** Purposive sampling is one in which the sample units are selected with definite purpose in view. For example, if we want to give the picture that the standard of living has increased in the city of New Delhi, we may take individuals in the sample from rich and posh localities like Defence Colony, South Extension, Golf Links, Jor Bagh, Chanakyapuri, Greater Kailash etc. and ignore the localities where low income group and the middle class families live. This sampling suffers from the drawback of favouritism and nepotism and does not give a representative sample of the population.

**12-2-2 Random Sampling.** In this case the sample units are selected at random and the drawback of purposive sampling, viz., favouritism or subjective element, is completely overcome. A *random sample* is one in which each unit of population has an equal chance of being included in it.

Suppose we take a sample of size  $n$  from a finite population of size  $N$ . Then there are  ${}^N C_n$  possible samples. A sampling technique in which each of the  ${}^N C_n$  samples has an equal chance of being selected is known as *random sampling* and the sample obtained by this technique is termed as a *random sample*.

Proper care has to be taken to ensure that the selected sample is random. Human bias, which varies from individual to individual, is inherent in any sampling scheme administered by human beings. Fairly good random samples can be obtained by the use of *Tippet's random number tables* or by throwing of a dice, draw of a lottery, etc.

The simplest method, which is normally used, is the *lottery system* which is illustrated below by means of an example.

Suppose we want to select ' $r$ ' candidates out of  $n$ . We assign the numbers one to  $n$ , one number to each candidate and write these numbers (1 to  $n$ ) on  $n$  slips which are made as homogeneous as possible in shape, size, etc. These slips are then put in a bag and thoroughly shuffled and then ' $r$ ' slips are drawn one by one. The ' $r$ ' candidates corresponding to the numbers on the slips drawn, will constitute the random sample.

**Remark.** *Tippet's Random Numbers.* L.H.C. Tippet's random numbers tables consist of 10400 four-digit numbers, giving in all  $10400 \times 4$ , i.e., 41600 digits, taken from the British census reports. These tables have proved to be fairly random in character. Any page of the table is selected at random and the number in any row or column or diagonal selected at random may be taken to constitute the sample.

**12-2-3. Simple Sampling.** Simple sampling is random sampling in which each unit of the population has an equal chance, say  $p$ , of being included in the sample and that this probability is independent of the previous drawings. Thus a simple sample of size  $n$  from a population may be identified with a series of  $n$  independent trials with constant probability ' $p$ ' of success for each trial.

**Remark.** It may be pointed out that random sampling does not necessarily imply simple sampling though, obviously, the converse is true. For example, if

an urn contains ' $a$ ' white balls and ' $b$ ' black balls, the probability of drawing a white ball at the first draw is  $[a/(a+b)] = p_1$ , (say) and if this ball is not replaced the probability of getting a white ball in the second draw is  $[(a-1)(a+b-1)] = p_2 \neq p_1$ , the sampling is not simple. But since in the first draw each white ball has the same chance, viz.,  $a/(a+b)$ , of being drawn and in the second draw again each white ball has the same chance, viz.,  $(a-1)/(a+b-1)$ , of being drawn, the sampling is random. Hence in this case, the sampling, though random, is not simple. To ensure that sampling is simple, it must be done with replacement, if population is finite. However, in case of infinite population no replacement is necessary.

**12.2.4. Stratified Sampling.** Here the entire heterogeneous population is divided into a number of homogeneous groups, usually termed as *strata*, which differ from one another but each of these groups is homogeneous within itself. Then units are sampled at random from each of these stratum, the sample size in each stratum varies according to the relative importance of the stratum in the population. The sample, which is the aggregate of the sampled units of each of the stratum, is termed as *stratified sample* and the technique of drawing this sample is known as *stratified sampling*. Such a sample is by far the best and can safely be considered as representative of the population from which it has been drawn.

**12.3. Parameter and Statistic.** In order to avoid verbal confusion with the statistical constants of the population, viz., mean ( $\mu$ ), variance  $\sigma^2$ , etc., which are usually referred to as *parameters*, statistical measures computed from the sample observations alone, e.g., mean ( $\bar{x}$ ), variance ( $s^2$ ), etc., have been termed by Professor R.A. Fisher as *statistics*.

In practice, parameter values are not known and the estimates based on the sample values are generally used. Thus statistic which may be regarded as an estimate of parameter, obtained from the sample, is a function of the sample values only. It may be pointed out that a statistic, as it is based on sample values and as there are multiple choices of the samples that can be drawn from a population, varies from sample to sample. The determination or the characterisation of the variation (in the values of the statistic obtained from different samples) that may be attributed to chance or fluctuations of sampling is one of the fundamental problems of the sampling theory.

**Remarks 1.** Now onwards,  $\mu$  and  $\sigma^2$  will refer to the population mean and variance respectively while the sample mean and variance will be denoted by  $\bar{x}$  and  $s^2$  respectively.

**2. Unbiased Estimate.** A statistic  $t = t(x_1, x_2, \dots, x_n)$ , a function of the sample values  $x_1, x_2, \dots, x_n$  is an unbiased estimate of population parameter  $\theta$ , if  $E(t) = \theta$ . In other words, if

$$E(\text{Statistic}) = \text{Parameter}, \quad \dots(12-1)$$

then statistic is said to be an unbiased estimate of the parameter.

**12.3.1. Sampling Distribution of a Statistic.** If we draw a sample of size  $n$  from a given finite population of size  $N$ , then the total number of possible samples is :

$${}^N C_n = \frac{N!}{n!(N-n)!} = k, \text{ (say).}$$

For each of these  $k$  samples we can compute some statistic  $t = t(x_1; x_2, \dots, x_n)$ , in particular the mean  $\bar{x}$ , the variance  $s^2$ , etc., as given below :

Sample Number	Statistics		
	$t$	$\bar{x}$	$s^2$
1	$t_1$	$\bar{x}_1$	$s_1^2$
2	$t_2$	$\bar{x}_2$	$s_2^2$
3	$t_3$	$\bar{x}_3$	$s_3^2$
:	:	:	:
:	:	:	:
$k$	$t_k$	$\bar{x}_k$	$s_k^2$

The set of the values of the statistic so obtained, one for each sample, constitutes what is called the *sampling distribution* of the statistic. For example, the values  $t_1, t_2, t_3, \dots, t_k$  determine the sampling distribution of the statistic  $t$ . In other words, statistic  $t$  may be regarded as a random variable which can take the values  $t_1, t_2, t_3, \dots, t_k$  and we can compute the various statistical constants like mean, variance, skewness, kurtosis etc., for its distribution. For example, the mean and variance of the sampling distribution of the statistic  $t$  are given by :

$$\bar{t} = \frac{1}{k} (t_1 + t_2 + \dots + t_k) = \frac{1}{k} \sum_{i=1}^k t_i$$

$$\begin{aligned}\text{Var}(t) &= \frac{1}{k} [(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \dots + (t_k - \bar{t})^2] \\ &= \frac{1}{k} \sum_{i=1}^k (t_i - \bar{t})^2\end{aligned}$$

**12.3.2. Standard Error.** The standard deviation of the sampling distribution of a statistic is known as its *Standard Error*, abbreviated as S.E. The standard errors of some of the well known statistics, for large samples, are given below, where  $n$  is the sample size,  $\sigma^2$  the population variance, and  $P$  the population proportion, and  $Q = 1 - P$ ,  $n_1$  and  $n_2$  represent the sizes of two independent random samples respectively drawn from the given population(s).

S.No.	Statistic	Standard Error
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion ' $p$ '	$\sqrt{PQ/n}$
3.	Sample s.d. : $s$	$\sqrt{\sigma^2/2n}$
4.	Sample variance : $s^2$	$\sigma^2\sqrt{2/n}$
5.	Sample quartiles	1.36263 $\sigma/\sqrt{n}$
6.	Sample median	1.25331 $\sigma/\sqrt{n}$

7.	Sample correlation coefficient ( $r$ )	$(1 - \rho^2)/\sqrt{n}$ , $\rho$ being the population correlation coefficient
8.	Sample moment $\mu_3$	$\sigma^3 \sqrt{96/n}$
9.	Sample moment $\mu_4$	$\sigma^4 \sqrt{96/n}$
10.	Sample coefficient of variation ( $v$ )	$\frac{v}{\sqrt{2n}} \sqrt{1 + \frac{2v^3}{10^4}} \approx \frac{v}{\sqrt{2n}}$
11.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
12.	Difference of two sample s.d.'s : $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
13.	Difference of two sample proportions ( $p_1 - p_2$ )	$\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}$

**Remark on the Utility of Standard Error.** S.E. plays a very important role in the large sample theory and forms the basis of the testing of hypothesis. If  $t$  is any statistic, then for large samples

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0, 1) \quad (\text{c.f. } \S \text{ 12.9})$$

$$\Rightarrow Z = \frac{t - E(t)}{\text{S.E.}(t)} \sim N(0, 1), \text{ for large samples.}$$

Thus, if the discrepancy between the observed and the expected (hypothetical) value of a statistic is greater than  $z_\alpha$  (c.f.  $\S$  12.7.2) times its S.E., the null hypothesis is rejected at  $\alpha$  level of significance. Similarly, if

$$|t - E(t)| \leq z_\alpha \times \text{S.E.}(t),$$

the deviation is not regarded significant at 5% level of significance. In other words, the deviation,  $t - E(t)$ , could have arisen due to fluctuations of sampling and the data do not provide us any evidence against the null hypothesis which may, therefore, be accepted at  $\alpha$  level of significance. [For details see  $\S$  12.7.3]

(i) The magnitude of the standard error gives an index of the precision of the estimate of the parameter. The reciprocal of the standard error is taken as the measure of reliability or precision of the statistic.

$$\text{S.E.}(p) = \sqrt{PQ/n} \quad [\text{c.f. } (4b) \text{ } \S \text{ 12.9.1}]$$

$$\text{and} \quad \text{S.E.}(\bar{x}) = \sigma/\sqrt{n} \quad [\text{c.f. } \S \text{ 12.2}]$$

In other words, the standard errors of  $p$  and  $\bar{x}$  vary inversely as the square root of the sample size. Thus in order to double the precision, which amounts to reducing the standard error to one half, the sample size has to be increased four times.

(ii) S.E. enables us to determine the probable limits within which the population parameter may be expected to lie. For example, the probable limits for population proportion  $P$  are given by

$$p \pm 3\sqrt{pq/n}$$

(c.f. Remark § 12.9.1)

**Remark.** S.E. of a statistic may be reduced by increasing the sample size but this results in corresponding increase in cost, labour and time, etc.

**12.4. Tests of Significance.** A very important aspect of the sampling theory is the study of the *tests of significance*, which enable us to decide on the basis of the sample results, if

(i) the deviation between the observed sample statistic and the hypothetical parameter value, or

(ii) the deviation between two independent sample statistics;

is significant or might be attributed to chance or the fluctuations of sampling.

Since, for large  $n$ , almost all the distributions, e.g., Binomial, Poisson, Negative binomial, Hypergeometric (c.f. Chapter 7),  $t$ ,  $F$  (Chapter 14), Chi-square (Chapter 13), can be approximated very closely by a normal probability curve, we use the *Normal Test of Significance* (c.f. § 12.9) for large samples. Some of the well known tests of significance for studying such differences for small samples are *t-test*, *F-test* and Fisher's *z-transformation*.

**12.5. Null Hypothesis.** The technique of randomisation used for the selection of sample units makes the test of significance valid for us. For applying the test of significance we first set up a hypothesis—a definite statement about the population parameter. Such a hypothesis, which is usually a hypothesis of no difference, is called *null hypothesis* and is usually denoted by  $H_0$ . According to Prof. R.A. Fisher, *null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true*.

For example, in case of a single statistic,  $H_0$  will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics,  $H_0$  will be that the sample statistics do not differ significantly.

Having set up the null hypothesis we compute the probability  $P$  that the deviation between the observed sample statistic and the hypothetical parameter value might have occurred due to fluctuations of sampling (c.f. § 12.7). If the deviation comes out to be significant (as measured by a test of significance), null hypothesis is refuted or rejected at the particular level of significance adopted (c.f. § 12.7) and if the deviation is not significant, null hypothesis may be retained at that level.

**12.5.1. Alternative Hypothesis.** Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by  $H_1$ . For example, if we want to test the null hypothesis that the population has a specified mean  $\mu_0$ , (say), i.e.,  $H_0 : \mu = \mu_0$ , then the alternative hypothesis could be

(i)  $H_1 : \mu \neq \mu_0$  (i.e.,  $\mu > \mu_0$  or  $\mu < \mu_0$ )

(ii)  $H_1 : \mu > \mu_0$

(iii)  $H_1 : \mu < \mu_0$

The alternative hypothesis in (i) is known as a *two tailed alternative* and the alternatives in (ii) and (iii) are known as *right tailed* and *left-tailed alternatives* respectively. The setting of alternative hypothesis is very important since it

enables us to decide whether we have to use a single-tailed (right or left) or two-tailed test [c.f. § 12-7-1].

**12-6. Errors in Sampling.** The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. In practice we decide to accept or reject the lot after examining a sample from it. As such we are liable to commit the following two types of errors :

**Type I Error** : *Reject  $H_0$  when it is true.*

**Type II Error** : *Accept  $H_0$  when it is wrong, i.e., accept  $H_0$  when  $H_1$  is true.*

If we write,

$$\left. \begin{aligned} P\{\text{Reject } H_0 \text{ when it is true}\} &= P\{\text{Reject } H_0 | H_0\} = \alpha \\ \text{and } P\{\text{Accept } H_0 \text{ when it is wrong}\} &= P\{\text{Accept } H_0 | H_1\} = \beta \end{aligned} \right\} \dots(12-2)$$

then  $\alpha$  and  $\beta$  are called the *sizes of type I error and type II error*, respectively.

In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad.

Thus

$$\left. \begin{aligned} P\{\text{Reject a lot when it is good}\} &= \alpha \\ \text{and } P\{\text{Accept a lot when it is bad}\} &= \beta \end{aligned} \right\} \dots(12-2a)$$

where  $\alpha$  and  $\beta$  are referred to as *Producer's risk* and *Consumer's risk*, respectively.

**12-7. Critical Region and Level of Significance.** A region (corresponding to a statistic  $t$ ) in the sample space  $S$  which amounts to rejection of  $H_0$  is termed as *critical region* or *region of rejection*. If  $\omega$  is the critical region and if  $t = t(x_1, x_2, \dots, x_n)$  is the value of the statistic based on a random sample of size  $n$ , then

$$P(t \in \omega | H_0) = \alpha, \quad P(t \in \bar{\omega} | H_1) = \beta \quad (12-2b)$$

where  $\bar{\omega}$ , the complementary set of  $\omega$ , is called the *acceptance region*.

We have  $\omega \cup \bar{\omega} = S$  and  $\omega \cap \bar{\omega} = \emptyset$

The probability ' $\alpha$ ' that a random value of the statistic  $t$  belongs to the critical region is known as the *level of significance*. In other words, level of significance is the size of the type I error (or the maximum producer's risk). The levels of significance usually employed in testing of hypothesis are 5% and 1%. The level of significance is always fixed in advance before collecting the sample information.

**12-7-1. One tailed and Two Tailed Tests.** In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

A test of any statistical hypothesis where the alternative hypothesis is one tailed (right tailed or left tailed) is called a *one tailed test*. For example, a test for testing the mean of a population

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis :

$$H_1 : \mu > \mu_0 \text{ (Right tailed)} \text{ or } H_1 : \mu < \mu_0 \text{ (Left tailed),}$$

is a *single tailed test*: In the right tailed test ( $H_1 : \mu_1 > \mu_0$ ), the critical region lies entirely in the right tail of the sampling distribution of  $\bar{x}$ , while for the left tail test ( $H_1 : \mu_1 < \mu_0$ ), the critical region is entirely in the left tail of the 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.

In a particular problem, whether one tailed or two tailed test is to be applied depends entirely on the nature of the alternative hypothesis. If the alternative hypothesis is two-tailed we apply two-tailed test and if alternative hypothesis is one-tailed, we apply one tailed test.

For example, suppose that there are two population brands of bulbs, one manufactured by standard process (with mean life  $\mu_1$ ) and the other manufactured by some new technique (with mean life  $\mu_2$ ). If we want to test if the bulbs differ significantly, then our null hypothesis is  $H_0 : \mu_1 = \mu_2$  and alternative will be  $H_1 : \mu_1 \neq \mu_2$ , thus giving us a two-tailed test. However, if we want to test if the bulbs produced by new process have higher average life than those produced by standard process, then we have

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 < \mu_2,$$

thus giving us a left-tail test. Similarly, for testing if the product of new process is inferior to that of standard process, then we have :

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 > \mu_2,$$

thus giving us a right-tail test. Thus, the decision about applying a two-tail test or a single-tail (right or left) test will depend on the problem under study.

**12.7.2. Critical Values or Significant Values.** The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the *critical value* or *significant value*. It depends upon :

- (i) The level of significance used, and
- (ii) The alternative hypothesis, whether it is two-tailed or single-tailed.

As has been pointed out earlier, for large samples, the standardised variable corresponding to the statistic  $t$  viz. :

$$Z = \frac{t - E(t)}{S.E.(t)} \sim N(0, 1), \quad \dots (*)$$

asymptotically as  $n \rightarrow \infty$ . The value of  $Z$  given by (\*) under the null hypothesis is known as *test statistic*. The critical value of the test statistic at level of significance  $\alpha$  for a two-tailed test is given by  $z_\alpha$  where  $z_\alpha$  is determined by the equation

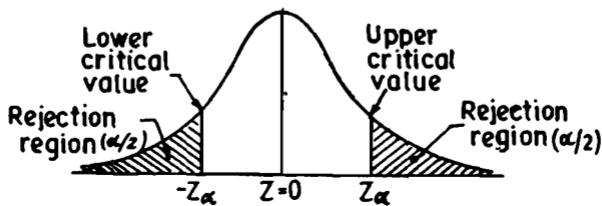
$$P(|Z| > z_\alpha) = \alpha \quad \dots (12.2c)$$

i.e.,  $z_\alpha$  is the value so that the total area of the critical region on both tails is  $\alpha$ . Since normal probability curve is a symmetrical curve, from (12.2c), we get

$$\begin{aligned} P(Z > z_\alpha) + P(Z < -z_\alpha) &= \alpha && [\text{By symmetry}] \\ \Rightarrow P(Z > z_\alpha) + P(Z > z_\alpha) &= \alpha \\ \Rightarrow 2P(Z > z_\alpha) &= \alpha \\ \Rightarrow P(Z > z_\alpha) &= \frac{\alpha}{2} \end{aligned}$$

i.e., the area of each tail is  $\alpha/2$ . Thus  $z_\alpha$  is the value such that area to the right of  $z_\alpha$  is  $\alpha/2$  and to the left of  $-z_\alpha$  is  $\alpha/2$ , as shown in the following diagram.

TWO-TAILED TEST  
(Level of Significance ' $\alpha$ ')



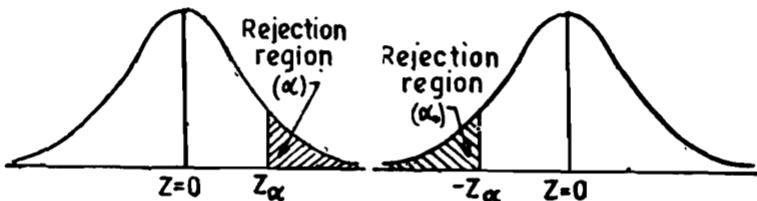
In case of single-tail alternative, the critical value  $z_\alpha$  is determined so that total area to the right of it (for right-tailed test) is  $\alpha$  and for left-tailed test the total area to the left of  $-z_\alpha$  is  $\alpha$  (See diagrams below), i.e.,

$$\text{For Right-tail Test : } P(Z > z_\alpha) = \alpha \quad \dots(12.2d)$$

$$\text{For Left-tail Test : } P(Z < -z_\alpha) = \alpha \quad \dots(12.2e)$$

RIGHT-TAILED TEST  
(Level of Significance ' $\alpha$ ')

LEFT-TAILED TEST  
(Level of Significance ' $\alpha'$ )



Thus the significant or critical value of  $Z$  for a single-tailed test (left or right) at level of significance ' $\alpha$ ' is same as the critical value of  $Z$  for a two-tailed test at level of significance ' $2\alpha$ '.

We give on page 12-10, the critical values of  $Z$  at commonly used levels of significance for both two-tailed and single-tailed tests. These values have been obtained from equations (12.2c), (12.2d) and (12.2e), on using the Normal Probability Tables as explained in § 12-8.

CRITICAL VALUES ( $z_\alpha$ ) OF Z

Critical Values ( $z_\alpha$ )	Level of significance ( $\alpha$ )		
	1%	5%	10%
Two-tailed test	$ Z_\alpha  = 2.58$	$ Z_\alpha  = 1.96$	$ Z_\alpha  = 1.645$
Right-tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left-tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

**Remark.** If  $n$  is small, then the sampling distribution of the test statistic  $Z$  will not be normal and in that case we can't use the above significant values, which have been obtained from normal probability curves. In this case, viz.,  $n$  small, (usually less than 30), we use the significant values based on the exact sampling distribution of the statistic  $Z$ , [defined in (\*), § 12-7-2], which turns out to be  $t$ ,  $F$ , or  $\chi^2$  [see Chapters 13, 14]. These significant values have been tabulated for different values of  $n$  and  $\alpha$  and are given in the Appendix at the end of the book.

**12-7-3. Procedure for Testing of Hypothesis.** We now summarise below the various steps in testing of a statistical hypothesis in a systematic manner.

1. *Null Hypothesis.* Set up the Null Hypothesis  $H_0$  (see § 12-5, page 12-6).

2. *Alternative Hypothesis.* Set up the Alternative Hypothesis  $H_1$ . This will enable us to decide whether we have to use a single-tailed (right or left) test or two-tailed test.

3. *Level of Significance.* Choose the appropriate level of significance ( $\alpha$ ) depending on the reliability of the estimates and permissible risk. This is to be decided before sample is drawn, i.e.,  $\alpha$  is fixed in advance.

4. *Test Statistic (or Test Criterion).* Compute the test statistic

$$Z = \frac{t - E(t)}{S.E.(t)}$$

under the null hypothesis.

5. *Conclusion.* We compare  $z$  the computed value of  $Z$  in step 4 with the significant value (tabulated value)  $z_\alpha$ , at the given level of significance, ' $\alpha$ '.

If  $|Z| < z_\alpha$ , i.e., if the calculated value of  $Z$  (in modulus value) is less than  $z_\alpha$  we say it is not significant. By this we mean that the difference  $t - E(t)$  is just due to fluctuations of sampling and the sample data do not provide us sufficient evidence against the null hypothesis which may therefore, be accepted.

If  $|Z| > z_\alpha$ , i.e., if the computed value of test statistic is greater than the critical or significant value, then we say that it is significant and the null hypothesis is rejected at level of significance  $\alpha$  i.e., with confidence coefficient  $(1 - \alpha)$ .

**12-8. Test of Significance for Large Samples.** In this section we

will discuss the tests of significance when samples are large. We have seen that for large values of  $n$ , the number of trials, almost all the distributions, e.g., binomial, Poisson, negative binomial, etc., are very closely approximated by normal distribution. Thus in this case we apply the *normal test*, which is based upon the following fundamental property (*area property*) of the normal probability curve.

$$\text{If } X \sim N(\mu, \sigma^2), \text{ then } Z = \frac{X - \mu}{\sigma} = \frac{X - E(X)}{\sqrt{V(X)}} \sim N(0, 1)$$

Thus from the normal probability tables, we have

$$\begin{aligned} P(-3 \leq Z \leq 3) &= 0.9973, \text{ i.e., } P(|Z| \leq 3) = 0.9973 \\ \Rightarrow P(|Z| > 3) &= 1 - P(|Z| \leq 3) = 0.0027 \end{aligned} \quad \dots(12.3)$$

i.e., in all probability we should expect a standard normal variate to lie between  $\pm 3$ .

Also from the normal probability tables, we get

$$P(-1.96 \leq Z \leq 1.96) = 0.95 \text{ i.e., } P(|Z| \leq 1.96) = 0.95$$

$$\Rightarrow P(|Z| > 1.96) = 1 - 0.95 = 0.05 \quad \dots(12.3a)$$

$$\text{and } P(|Z| \leq 2.58) = 0.99$$

$$\Rightarrow P(|Z| > 2.58) = 0.01 \quad \dots(12.3b)$$

Thus the significant values of  $Z$  at 5% and 1% level of significance for a two tailed test are 1.96 and 2.58 respectively.

Thus the steps to be used in the normal test are as follows :

(i) Compute the test statistic  $Z$  under  $H_0$ .

(ii) If  $|Z| > 3$ ,  $H_0$  is always rejected.

(iii) If  $|Z| \leq 3$ , we test its significance at certain level of significance, usually at 5% and sometimes at 1% level of significance. Thus, for a two-tailed test if  $|Z| > 1.96$ ,  $H_0$  is rejected at 5% level of significance.

Similarly if  $|Z| > 2.58$ ,  $H_0$  is contradicted at 1% level of significance and if  $|Z| \leq 2.58$ ,  $H_0$  may be accepted at 1% level of significance.

From the normal probability tables, we have :

$$\begin{aligned} P(Z > 1.645) &= 0.5 - P(0 \leq Z \leq 1.645) \\ &= 0.5 - 0.45 \\ &= 0.05 \end{aligned}$$

$$\begin{aligned} P(Z > 2.33) &= 0.5 - P(0 \leq Z \leq 2.33) \\ &= 0.5 - 0.49 \\ &= 0.01 \end{aligned}$$

Hence for a single-tail test (Right-tail or Left-tail) we compare the computed value of  $|Z|$  with 1.645 (at 5% level) and 2.33 (at 1% level) and accept or reject  $H_0$  accordingly.

**Important Remark.** In the theoretical discussion that follows in the next sections, the samples under consideration are supposed to be large. For practical purposes, sample may be regarded as large if  $n > 30$ .

**12.9. Sampling of Attributes.** Here we shall consider sampling from a population which is divided into two mutually exclusive and collectively

exhaustive classes—one class possessing a particular attribute, say  $A$ , and the other class not possessing that attribute, and then note down the number of persons in the sample of size  $n$ , possessing that attribute. The presence of an attribute in sampled unit may be termed as success and its absence as failure. In this case a sample of  $n$  observations is identified with that of a series of  $n$  independent Bernoulli trials with constant probability  $P$  of success for each trial. Then the probability of  $x$  successes in  $n$  trials, as given by the binomial probability distribution is

$$p(x) = {}^n C_x P^x Q^{n-x}; x = 0, 1, 2, \dots, n.$$

**12.9.1. Test for Single Proportion.** If  $X$  is the number of successes in  $n$  independent trials with constant probability  $P$  of success for each trial (c.f. § 7.2-1)

$$E(X) = nP \text{ and } V(X) = nPQ,$$

where  $Q = 1 - P$ , is the probability of failure.

It has been proved that for large  $n$ , the binomial distribution tends to normal distribution. Hence for large  $n$ ,  $X \sim N(nP, nPQ)$  i.e.,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0, 1) \quad \dots(12.4)$$

and we can apply the normal test.

**Remarks 1.** In a sample of size  $n$ , let  $X$  be the number of persons possessing the given attribute. Then

Observed proportion of successes  $= X/n = p$ , (say).

$$\begin{aligned} \therefore E(p) &= E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} nP = P \\ \Rightarrow E(p) &= P \end{aligned} \quad \dots(12.4a)$$

Thus the sample proportion ' $p$ ' gives an unbiased estimate of the population proportion  $P$ .

$$\text{Also } V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} nPQ = \frac{PQ}{n}$$

$$\therefore S.E.(p) = \sqrt{PQ/n} \quad \dots(12.4b)$$

Since  $X$  and consequently  $X/n$  is asymptotically normal for large  $n$ , the normal test for the proportion of successes becomes

$$Z = \frac{p - E(p)}{S.E.(p)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1) \quad \dots(12.4c)$$

**2.** If we have sampling from a finite population of size  $N$ , then

$$S.E.(p) = \sqrt{\left(\frac{N-n}{N-1}\right) \cdot \frac{PQ}{n}} \quad \dots(12.4d)$$

**3.** Since the probable limits for a normal variate  $X$  are  $E(X) \pm 3 \sqrt{V(X)}$ , the probable limits for the observed proportion of successes are :

$$E(p) \pm 3 S.E.(p), \text{ i.e., } P \pm 3 \sqrt{PQ/n}.$$

If  $P$  is not known then taking  $p$  (the sample proportion) as an estimate of  $P$ , the probable limits for the proportion in the population are :

$$p \pm 3 \sqrt{pq/n} \quad \dots(12.4e)$$

However, the limits for  $P$  at level of significance  $\alpha$  are given by :

$$p \pm z_\alpha \sqrt{pq/n}, \quad \dots(12.4f)$$

where  $z_\alpha$  is the significant value of  $Z$  at level of significance  $\alpha$ .

In particular 95% confidence limits for  $P$  are given by :

$$p \pm 1.96 \sqrt{pq/n}, \quad \dots(12.4g)$$

and 99% confidence limits for  $P$  are given by

$$p \pm 2.58 \sqrt{pq/n} \quad \dots(12.4h)$$

**Example 12.1.** A dice is thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the dice cannot be regarded as an unbiased one and find the limits between which the probability of a throw of 3 or 4 lies.

**Solution.** If the coming of 3 or 4 is called a success, then in usual notations we are given

$$n = 9,000; X = \text{Number of successes} = 3,240$$

Under the null hypothesis ( $H_0$ ) that the dice is an unbiased one, we get

$$P = \text{Probability of success} = \text{Probability of getting a 3 or 4} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Alternative hypothesis,  $H_1 : p \neq \frac{1}{3}$ , (i.e., dice is biased).

We have  $Z = \frac{X - nP}{\sqrt{nQP}} \sim N(0, 1)$ , since  $n$  is large.

$$\text{Now } Z = \frac{3240 - 9000 \times 1/3}{\sqrt{9000 \times (1/3) \times (2/3)}} = \frac{240}{\sqrt{2000}} = \frac{240}{44.73} = 5.36$$

Since  $|Z| > 3$ ,  $H_0$  is rejected and we conclude that the dice is almost certainly biased.

Since dice is not unbiased,  $P \neq \frac{1}{3}$ . The probable limits for ' $P$ ' are given by :

$$\hat{P} \pm 3 \sqrt{\hat{P}\hat{Q}/n} = p \pm 3 \sqrt{pq/n},$$

$$\text{where } \hat{P} = p = \frac{3240}{9000} = 0.36 \text{ and } \hat{Q} = q = 1 - p = 0.64.$$

Hence the probable limits for the population proportion of successes may be taken as

$$\begin{aligned} \hat{P} \pm 3 \sqrt{\hat{P}\hat{Q}/n} &= 0.36 \pm 3 \sqrt{\frac{0.36 \times 0.64}{9000}} = 0.36 \pm 3 \times \frac{0.6 \times 0.8}{30 \sqrt{10}} \\ &= 0.360 \pm 0.015 = 0.345 \text{ and } 0.375. \end{aligned}$$

Hence the probability of getting 3 or 4 almost certainly lies between 0.345 and 0.375.

**Example 12.2.** A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the S.E. of the

proportion of bad ones in a sample of this size is 0.015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.

**Solution.** Here we are given  $n = 500$

$X$  = Number of bad pineapples in the sample = 65

$$p = \text{Proportion of bad pineapples in the sample} = \frac{65}{500} = 0.13$$

$$\therefore q = 1 - p = 0.87$$

Since  $P$ , the proportion of bad pineapples in the consignment is not known, we may take (as in the last example)

$$\hat{P} = p = 0.13, \quad \hat{Q} = q = 0.87$$

$$\text{S.E. of proportion} = \sqrt{\frac{\hat{P}\hat{Q}}{n}} = \sqrt{0.13 \times 0.87/500} = 0.015$$

Thus, the limits for the proportion of bad pineapples in the consignment are :

$$\hat{P} \pm 3 \sqrt{\frac{\hat{P}\hat{Q}}{n}} = 0.130 \pm 3 \times 0.015 = 0.130 \pm 0.045 = (0.085, 0.175)$$

Hence the percentage of bad pineapples in the consignment lies almost certainly between 8.5 and 17.5.

**Example 12-3.** A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain the 98% confidence limits for the percentage number of bad apples in the consignment.

$$\left[ \int_0^{2.33} \phi(t) dt = 0.49 \text{ nearly} \right]$$

**Solution.** We have :

$$p = \text{Proportion of bad apples in the sample} = \frac{60}{500} = 0.12$$

Since the significant value of  $Z$  at 98% confidence coefficient (level of significance 2%) is given to be 2.33, 98% confidence limits for population proportion are :

$$\begin{aligned} p \pm 2.33 \sqrt{pq/n} &= 0.12 \pm 2.33 \sqrt{0.12 \times 0.88/500} \\ &= 0.12 \pm 2.33 \times \sqrt{0.0002112} = 0.12 \pm 2.33 \times 0.01453 \\ &= 0.12000 \pm 0.03385 = (0.08615, 0.15385) \end{aligned}$$

Hence 98% confidence limits for percentage of bad apples in the consignment are (8.61, 15.38).

**Example 12-4.** In a sample of 1,000 people in Maharashtra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this State at 1% level of significance?

**Solution.** In the usual notations we are given  $n = 1,000$

$X$  = Number of rice eaters = 540

$$\therefore p = \text{Sample proportion of rice eaters} = \frac{X}{n} = \frac{540}{1000} = 0.54$$

*Null Hypothesis,  $H_0$*  : Both rice and wheat are equally popular in the State so that

$$\begin{aligned} P &= \text{Population proportion of rice eaters in Maharashtra} = 0.5 \\ \Rightarrow Q &= 1 - P = 0.5 \end{aligned}$$

*Alternative Hypothesis,  $H_1$*  :  $P \neq 0.5$  (two-tailed alternative).

*Test Statistic.* Under  $H_0$ , the test statistic is

$$Z = \frac{P - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since } n \text{ is large).}$$

$$\text{Now } Z = \frac{0.54 - 0.50}{\sqrt{0.5 \times 0.5/1000}} = \frac{0.04}{0.0138} = 2.532$$

*Conclusion.* The significant or critical value of  $Z$  at 1% level of significance for two-tailed test is 2.58. Since computed  $Z = 2.532$  is less than 2.58, it is not significant at 1% level of significance. Hence the null hypothesis is accepted and we may conclude that rice and wheat are equally popular in Maharashtra State.

**Example 12.5.** Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more, at 5% level. (Use Large Sample Test.)

[Patna Univ. B.Sc. (Hons.), 1992; Bombay Univ. B.Sc. 1987]

**Solution.** In the usual notations, we are given  $n = 20$ .

$X$  = Number of persons who survived after attack by a disease = 18

$$p = \text{Proportion of persons survived in the sample} = \frac{18}{20} = 0.90$$

*Null Hypothesis,  $H_0$*  :  $P = 0.85$ , i.e., the proportion of persons survived after attack by a disease in the lot is 85%.

*Alternative Hypothesis,  $H_1$*  :  $P > 0.85$  (Right-tail alternative).

*Test Statistic.* Under  $H_0$ , the test statistic is :

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since sample is large).}$$

$$\text{Now } Z = \frac{0.90 - 0.85}{\sqrt{0.85 \times 0.15/20}} = \frac{0.05}{0.079} = 0.633$$

*Conclusion.* Since the alternative hypothesis is one-sided (right-tailed), we shall apply right-tailed test for testing significance of  $Z$ . The significant value of  $Z$  at 5% level of significance for right-tail test is + 1.645. Since computed value of  $Z = 0.633$  is less than 1.645, it is not significant and we may accept the null hypothesis at 5% level of significance.

**12.9.2. Test of Significance for Difference of Proportions.** Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say  $A$ , among their members. Let  $X_1, X_2$  be the number of persons possessing the given attribute  $A$  in random samples of sizes  $n_1$  and  $n_2$  from the two populations respectively. Then sample proportions are given by

$$p_1 = X_1/n_1 \text{ and } p_2 = X_2/n_2$$

If  $P_1$  and  $P_2$  are the population proportions, then

$$E(p_1) = P_1, E(p_2) = P_2 \quad [\text{c.f. Equation (12-4a)}]$$

and

$$V(p_1) = \frac{P_1 Q_1}{n_1} \text{ and } V(p_2) = \frac{P_2 Q_2}{n_2}$$

Since for large samples,  $p_1$  and  $p_2$  are asymptotically normally distributed,  $(p_1 - p_2)$  is also normally distributed. Then the standard variable corresponding to the difference  $(p_1 - p_2)$  is given by

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0, 1)$$

Under the *null hypothesis*  $H_0 : P_1 = P_2$ , i.e., there is no significant difference between the sample proportions, we have

$$E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2 = 0 \quad (\text{Under } H_0)$$

Also  $V(p_1 - p_2) = V(p_1) + V(p_2)$ ,

the covariance term  $\text{Cov}(p_1, p_2)$  vanishes, since sample proportions are independent.

$$\therefore V(p_1 - p_2) = \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2} = PQ \left( \frac{1}{n_1} + \frac{1}{n_2} \right),$$

since under  $H_0 : P_1 = P_2 = P$ , (say), and  $Q_1 = Q_2 = Q$ .

Hence under  $H_0 : P_1 = P_2$ , the test statistic for the difference of proportions becomes

$$Z = \frac{p_1 - p_2}{\sqrt{PQ \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1) \quad \dots(12-5)$$

In general, we do not have any information as to the proportion of  $A$ 's in the populations from which the samples have been taken. Under  $H_0 : P_1 = P_2 = P$ , (say), an unbiased estimate of the population proportion  $P$ , based on both the samples is given by

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} \quad \dots(12-5a)$$

The estimate is unbiased, since

$$\begin{aligned} E(\hat{P}) &= \frac{1}{n_1 + n_2} E[n_1 p_1 + n_2 p_2] = \frac{1}{n_1 + n_2} [n_1 E(p_1) + n_2 E(p_2)] \\ &= \frac{1}{n_1 + n_2} [n_1 P_1 + n_2 P_2] = P \quad [\because P_1 = P_2 = P, \text{ under } H_0] \end{aligned}$$

Thus (12-5) along with (12-5a) gives the required test statistic.

**Remarks 1.** Suppose we want to test the significance of the difference between  $p_1$  and  $p$ , where

$$p = \frac{(n_1 p_1 + n_2 p_2)}{(n_1 + n_2)}$$

gives a pooled estimate of the population proportion on the basis of both the samples. We have

$$V(p_1 - p) = V(p_1) + V(p) - 2 \operatorname{Cov}(p_1, p) \quad \dots(*)$$

Since  $p_1$  and  $p$  are not independent,  $\operatorname{Cov}(p_1, p) \neq 0$ .

$$\operatorname{Cov}(p_1, p) = E[(p_1 - E(p_1))(p - E(p))]$$

$$= E \left[ (p_1 - E(p_1)) \left\{ \frac{1}{n_1 + n_2} \{ n_1 p_1 + n_2 p_2 - E(n_1 p_1 + n_2 p_2) \} \right\} \right]$$

$$= \frac{1}{n_1 + n_2} E \left[ (p_1 - E(p_1)) \left\{ n_1(p_1 - E(p_1)) + n_2(p_2 - E(p_2)) \right\} \right]$$

$$= \frac{1}{n_1 + n_2} \left[ n_1 E \left\{ p_1 - E(p_1) \right\}^2 + n_2 E \left\{ (p_1 - E(p_1))(p_2 - E(p_2)) \right\} \right]$$

$$= \frac{1}{n_1 + n_2} \left[ n_1 V(p_1) + n_2 \operatorname{Cov}(p_1, p_2) \right]$$

$$= \frac{1}{n_1 + n_2} n_1 V(p_1), \quad [\because \operatorname{Cov}(p_1, p_2) = 0]$$

$$= \frac{n_1}{n_1 + n_2} \cdot \frac{pq}{n_1} = \frac{pq}{n_1 + n_2}$$

$$\text{Also } \operatorname{Var}(p) = \frac{1}{(n_1 + n_2)^2} E \left[ (n_1 p_1 + n_2 p_2) - E(n_1 p_1 + n_2 p_2) \right]^2 \\ = \frac{1}{(n_1 + n_2)^2} \left[ n_1^2 \operatorname{Var}(p_1) + n_2^2 \operatorname{Var}(p_2) \right].$$

covariance term vanishes since  $p_1$  and  $p_2$  are independent.

$$\therefore \operatorname{Var}(p) = \frac{1}{(n_1 + n_2)^2} \left[ n_1^2 \cdot \frac{pq}{n_1} + n_2^2 \cdot \frac{pq}{n_2} \right] \\ = \frac{pq}{n_1 + n_2}$$

Substituting in (\*) and simplifying, we shall get

$$V(p_1 - p) = \frac{pq}{n_1} + \frac{pq}{n_1 + n_2} - 2 \frac{pq}{n_1 + n_2} = pq \left[ \frac{n_2}{n_1(n_1 + n_2)} \right]$$

Thus, the test statistic in this case becomes

$$Z = \frac{p_1 - p}{\sqrt{\frac{n_2}{(n_1 + n_2)} \cdot \frac{pq}{n_1}}} \sim N(0, 1) \quad \dots(12.5b)$$

2. Suppose the population proportions  $P_1$  and  $P_2$  are given to be distinctly different, i.e.,  $P_1 \neq P_2$  and we want to test if the difference  $(P_1 - P_2)$  in population proportions is likely to be hidden in simple samples of sizes  $n_1$  and  $n_2$  from the two populations respectively.

We have seen that in the usual notations,

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\text{S.E.}(p_1 - p_2)} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} \sim N(0, 1)$$

Here sample proportions are not given. If we set up the *null hypothesis*  $H_0 : p_1 = p_2$ , i.e., the samples will not reveal the difference in the population proportions or in other words the difference in population proportions is likely to be hidden in sampling, the test statistic becomes

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} \sim N(0, 1) \quad \dots(12.5c)$$

**Example 12.6.** Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal, are same against that they are not, at 5% level. [Agra Univ. M.A., 1992]

**Solution.** Null Hypothesis  $H_0 : P_1 = P_2 = P$ , (say), i.e., there is no significant difference between the opinion of men and women as far as proposal of flyover is concerned.

Alternative Hypothesis,  $H_1 : P_1 \neq P_2$  (two-tailed).

We are given :

$$n_1 = 400, X_1 = \text{Number of men favouring the proposal} = 200$$

$$n_2 = 600, X_2 = \text{Number of women favouring the proposal} = 325$$

$$\therefore p_1 = \text{Proportion of men favouring the proposal in the sample}$$

$$= \frac{X_1}{n_1} = \frac{200}{400} = 0.5$$

$$p_2 = \text{Proportion of women favouring the proposal in the sample}$$

$$= \frac{X_2}{n_2} = \frac{325}{600} = 0.541$$

**Test Statistic.** Since samples are large, the test statistic under the Null-Hypothesis,  $H_0$  is :

$$Z = \frac{P_1 - P_2}{\sqrt{\hat{P} \hat{Q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1)$$

$$\text{where } \hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{200 + 325}{400 + 600} = \frac{525}{1000} = 0.525$$

$$\Rightarrow \hat{Q} = 1 - \hat{P} = 1 - 0.525 = 0.475$$

$$\therefore Z = \frac{0.500 - 0.541}{\sqrt{0.525 \times 0.475 \times \left( \frac{1}{400} + \frac{1}{600} \right)}}$$

$$= \frac{-0.041}{\sqrt{0.525 \times 0.475 \times (10/2,400)}} \\ = \frac{-0.041}{\sqrt{0.001039}} = \frac{-0.041}{0.0323} = -1.269$$

**Conclusion.** Since  $|Z| = 1.269$  which is less than 1.96, it is not significant at 5% level of significance. Hence  $H_0$  may be accepted at 5% level of significance and we may conclude that men and women do not differ significantly as regards proposal of flyover is concerned.

**Example 12-7.** A company has the head office at Calcutta and a branch at Bombay. The personnel director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Calcutta, 62% favoured the new plan. At Bombay out of a sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level?

**Solution.** In the usual notations, we are given :

$$n_1 = 500, p_1 = 0.62 \text{ and } n_2 = 400, p_2 = 1 - 0.41 = 0.59$$

**Null hypothesis,**  $H_0 : P_1 = P_2$ , i.e., there is no significant difference between the two groups in their attitude towards the new plan.

**Alternative hypothesis,**  $H_1 : P_1 \neq P_2$  (Two-tailed).

**Test Statistic.** Under  $H_0$ , the test statistic for large samples is :

$$Z = \frac{P_1 - P_2}{\text{S.E.}(P_1 - P_2)} = \frac{P_1 - P_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where  $\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{500 \times 0.62 + 400 \times 0.59}{500 + 400} = 0.607$

and  $\hat{Q} = 1 - \hat{P} = 0.393$

$$\therefore Z = \frac{0.62 - 0.59}{\sqrt{0.607 \times 0.393 \times \left(\frac{1}{500} + \frac{1}{400}\right)}} \\ = \frac{0.03}{\sqrt{0.00107}} = \frac{0.03}{0.0327} = 0.917.$$

**Critical region.** At 5% level of significance, the critical value of  $Z$  for a two-tailed test is 1.96. Thus the critical region consists of all values of  $Z \geq 1.96$  or  $Z \leq -1.96$ .

**Conclusion.** Since the calculated value of  $|Z| = 0.917$  is less than the critical value of  $Z$  (1.96), it is not significant at 5% level of significance. Hence the data do not provide us any evidence against the null hypothesis which may be accepted, and we conclude that there is no significant difference between the two groups in their attitude towards the new plan.

**Example 12-8.** Before an increase in excise duty on tea, 800 persons out of a sample of 1,000 persons were found to be tea drinkers. After an increase in

duty, 800 people were tea drinkers in a sample of 1,200 people. Using standard error of proportion, state whether there is a significant decrease in the consumption of tea after the increase in excise duty?

**Solution.** In the usual notations, we have  $n_1 = 1,000$ ;  $n_2 = 1,200$

$p_1$  = Sample proportion of tea drinkers before increase in excise duty

$$= \frac{800}{1000} = 0.80$$

$p_2$  = Sample proportion of tea drinkers after increase in excise duty

$$= \frac{800}{1200} = 0.67$$

**Null Hypothesis.**,  $H_0 : P_1 = P_2$ , i.e., there is no significant difference in the consumption of tea before and after the increase in excise duty.

**Alternative Hypothesis.**,  $H_1 : P_1 > P_2$  (Right-tailed alternative).

**Test Statistic.** Under the null hypothesis, the test statistic is

$$Z = \frac{P_1 - P_2}{\sqrt{\hat{P}\hat{Q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1) \quad (\text{Since samples are large})$$

where

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{16}{22}, \text{ and } \hat{Q} = 1 - \hat{P} = \frac{6}{22}$$

$$\therefore Z = \frac{0.80 - 0.67}{\sqrt{\frac{16}{22} \times \frac{6}{22} \times \left( \frac{1}{1000} + \frac{1}{1200} \right)}} \\ = \frac{0.13}{\sqrt{\frac{16}{22} \times \frac{6}{22} \times \frac{11}{6000}}} = \frac{0.13}{0.019} = 6.842$$

**Conclusion.** Since  $Z$  is much greater than 1.645 as well as 2.33 (since test is one-tailed), it is highly significant at both 5% and 1% levels of significance. Hence,

we reject the null hypothesis  $H_0$  and conclude that there is a significant decrease in the consumption of tea after increase in the excise duty.

**Example 12.9.** A cigarette manufacturing firm claims that its brand A of the cigarettes outsells its brand B by 8%. If it is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another random sample of 100 smokers prefer brand B, test whether the 8% difference is a valid claim. (Use 5% level of significance.)

**Solution.** We are given.:

$$n_1 = 200, X_1 = 42 \Rightarrow p_1 = \frac{X_1}{n_1} = \frac{42}{200} = 0.21$$

$$n_2 = 100, X_2 = 18 \Rightarrow p_2 = \frac{X_2}{n_2} = \frac{18}{100} = 0.18$$

We set up the Null Hypothesis that 8% difference in the sale of two brands of cigarettes is a valid claim, i.e.,  $H_0 : P_1 - P_2 = 0.08$ .

**Alternative Hypothesis :**  $H_1 : P_1 - P_2 \neq 0.08$  (Two-tailed).

Under  $H_0$ , the test statistic is (since samples are large)

$$Z = \frac{(P_1 - P_2) - (P_1 - P_2)}{\sqrt{\hat{P}\hat{Q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1)$$

where  $\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{42 + 18}{200 + 100} = \frac{60}{300} = 0.20 \Rightarrow \hat{Q} = 1 - \hat{P} = 0.80$

$$\therefore Z = \frac{(0.21 - 0.18) - (0.08)}{\sqrt{0.2 \times 0.8 \left( \frac{1}{200} + \frac{1}{100} \right)}} = \frac{-0.05}{\sqrt{0.16 \times 0.015}} \\ = \frac{-0.05}{\sqrt{0.0024}} = \frac{-0.05}{0.04899} = -1.02$$

Since  $|Z| = 1.02 < 1.96$ , it is not significant at 5% level of significance. Hence null hypothesis may be retained at 5% level of significance and we may conclude that a difference of 8% in the sale of two brands of cigarettes is a valid claim by the firm.

**Example 12.10.** On the basis of their total scores, 200 candidates of a civil service examination are divided into two groups, the upper 30 per cent and the remaining 70 per cent. Consider the first question of this examination. Among the first group, 40 had the correct answer, whereas among the second group, 80 had the correct answer. On the basis of these results, can one conclude that the first question is no good at discriminating ability of the type being examined here?

**Solution.** Here, we have

$$n = \text{Total number of candidates} = 200$$

$n_1$  = The number of candidates in the upper 30% group

$$= \frac{30}{100} \times 200 = 60$$

$n_2$  = The number of candidates in the remaining 70% group

$$= \frac{70}{100} \times 200 = 140$$

$X_1$  = The number of candidates, with correct answer in the first group = 40

$X_2$  = The number of candidates, with correct answer in the second group = 80

$$\therefore p_1 = \frac{X_1}{n_1} = \frac{40}{60} = 0.6666 \text{ and } p_2 = \frac{X_2}{n_2} = \frac{80}{140} = 0.5714$$

*Null Hypothesis*,  $H_0$  : There is no significant difference in the sample proportions, i.e.,  $P_1 = P_2$ , i.e., the first question is no good at discriminating the ability of the type being examined here.

*Alternative Hypothesis*,  $H_1 : P_1 \neq P_2$ .

*Test Statistic.* Under  $H_0$  the test statistic is :

$$\bar{Z} = \frac{P_1 - P_2}{\sqrt{\hat{P}\hat{Q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1) \quad (\text{since samples are large}).$$

where

$$\hat{P} = \frac{\bar{X}_1 + \bar{X}_2}{n_1 + n_2} = \frac{40 + 80}{60 + 140} = 0.6, \quad \hat{Q} = 1 - \hat{P} = 0.4$$

$$\therefore Z = \frac{0.6666 - 0.5714}{\sqrt{0.6 \times 0.4 \left( \frac{1}{60} + \frac{1}{140} \right)}} = \frac{0.0953}{0.0756} = 1.258$$

*Conclusion.* Since  $|Z| < 1.96$ , the data are consistent with the null hypothesis at 5% level of significance. Hence we conclude that the first question is not good enough to distinguish between the ability of the two groups of candidates.

**Example 12.11.** In a year there are 956 births in a town A, of which 52.5% were males, while in towns A and B combined, this proportion in a total of 1,406 births was 0.496. Is there any significant difference in the proportion of male births in the two towns?

**Solution.** We are given

$$n_1 = 956, \quad n_1 + n_2 = 1,406 \quad \text{or} \quad n_2 = 1,406 - 956 = 450$$

$$p_1 = \text{Proportion of males in the sample of town A} = 0.525.$$

Let  $p_2$  be the proportion of males in the sample (of size  $n_2$ ) of town B. Then

$\hat{P}$  = Proportion of males in both the samples combined.

$$= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0.496 \quad (\text{Given})$$

$$\therefore \frac{956 \times 0.525 + 450 \times p_2}{1,406} = 0.496$$

$$\Rightarrow p_2 = 0.434 \quad (\text{On simplification})$$

*Null Hypothesis*,  $H_0 : P_1 = P_2$ , i.e., there is no significant difference in the proportion of male births in the two towns A and B.

*Alternative Hypothesis*,  $H_1 : P_1 \neq P_2$  (two-tailed).

*Test Statistic.* Under  $H_0$ , the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P} \hat{Q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1) \quad (\text{Since samples are large})$$

$$\text{where } \hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0.496, \quad \hat{Q} = 1 - \hat{P} = 0.504$$

$$\therefore Z = \frac{0.525 - 0.434}{\sqrt{0.496 \times 0.504 \left( \frac{1}{956} + \frac{1}{450} \right)}} = \frac{0.091}{0.027} = 3.368$$

*Conclusion.* Since  $|Z| > 3$ , the null hypothesis is rejected, i.e., the data are inconsistent with the hypothesis  $P_1 = P_2$  and we conclude that there is significant difference in the proportion of male births in the towns A and B.

**Example 12-12.** In two large populations, there are 30 and 25 per cent respectively of blue-eyed people. Is this difference likely to be hidden in samples of 1,200 and 900 respectively from the two populations?

[Delhi Univ. B.Sc., 1992]

**Solution.** Here, we are given  $n_1 = 1200$ ,  $n_2 = 900$ .

$$\begin{aligned}P_1 &= \text{Proportion of blue-eyed people in the first population} \\&= 30\% = 0.30.\end{aligned}$$

$$\begin{aligned}P_2 &= \text{Proportion of blue-eyed people in the second population} \\&= 25\% = 0.25.\end{aligned}$$

$$\therefore Q_1 = 1 - P_1 = 0.70 \text{ and } Q_2 = 1 - P_2 = 0.75$$

We set up the *null hypothesis*  $H_0$  that  $p_1 = p_2$ , i.e., the sample proportions are equal, i.e., the difference in population proportions is likely to be hidden in sampling.

**Test Statistic.** Under  $H_0 : p_1 = p_2$ , the test statistic is :

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} \sim N(0, 1) \quad (\text{Since samples are large.})$$

$$\therefore |Z| = \frac{0.30 - 0.25}{\sqrt{\frac{0.3 \times 0.7}{1,200} + \frac{0.25 \times 0.75}{900}}} = \frac{0.05}{0.0195} = 2.56$$

**Conclusion.** Since  $|Z| > 1.96$ , the null hypothesis ( $p_1 = p_2$ ), is refuted at 5% level of significance and we conclude that the difference in population proportions is unlikely to be hidden in sampling. In other words, these samples will reveal the difference in the population proportions.

**Example 12-13.** In a random sample of 400 students of the university teaching departments, it was found that 300 students failed in the examination. In another random sample of 500 students of the affiliated colleges, the number of failures in the same examination was found to be 300. Find out whether the proportion of failures in the university teaching departments is significantly greater than the proportion of failures in the university teaching departments and affiliated colleges taken together.

**Solution.** Here we are given :  $n_1 = 400$ ,  $n_2 = 500$

$$p_1 = \frac{300}{400} = 0.75, \quad p_2 = \frac{300}{500} = 0.60$$

$$\therefore q_1 = 1 - p_1 = 1 - 0.75 = 0.25 \text{ and } q_2 = 1 - p_2 = 0.40$$

Here we set up the *null hypothesis*  $H_0$  that  $p_1$  and  $\hat{p}$ , where  $\hat{p}$  is the pooled estimate, i.e., proportion of failures in the university teaching departments and affiliated colleges taken together, do not differ significantly.

$$\text{S.E. of } (\hat{p} - p_1) = \sqrt{\frac{\hat{p} \hat{q}}{n_1 + n_2} \times \frac{n_2}{n_1}} \quad [\text{c.f. (12-5b) page 12-18}]$$

$$\text{where } \hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400 \times 0.75 + 500 \times 0.60}{400 + 500} = 0.67$$

$$\hat{q} = 1 - 0.67 = 0.33$$

$$\therefore \text{S.E. of } (\hat{p} - p_1) = \sqrt{\frac{0.67 \times 0.33}{400 + 500} \times \frac{500}{400}} = 0.018$$

*Test Statistic.* Under the null hypothesis  $H_0$ , the test statistic is :

$$Z = \frac{\hat{p} - p_1}{\text{S.E. of } (\hat{p} - p_1)} \sim N(0, 1) \quad (\text{Since samples are large.})$$

$$Z = \frac{0.67 - 0.33}{0.018} = \frac{0.15}{0.018} = 8.3$$

*Conclusion.* Since the calculated value of  $Z$  is much greater than 3, it is highly significant. Hence null hypothesis  $H_0$  is rejected and we conclude that there is significant difference between  $p_1$  and  $\hat{p}$ .

**Example 12-14.** If for one-half of  $n$  events, the chance of success is  $p$  and the chance of failure is  $q$ , while for the other half the chance of success is  $q$  and the chance of failure is  $p$ , show that the standard deviation of the number of successes is the same as if the chance of successes were  $p$  in all the cases, i.e.,  $\sqrt{npq}$  but that the mean of the number of successes is  $n/2$  and not  $np$ .

**Solution.** Let  $X_1$  and  $X_2$  denote the number of successes in the first half and the second half of  $n$  events respectively. Then according to the given conditions, we have

$$\left. \begin{array}{l} E(X_1) = \frac{n}{2} p \\ V(X_1) = \frac{n}{2} pq \end{array} \right\} \text{and} \left. \begin{array}{l} E(X_2) = \frac{n}{2} q \\ V(X_2) = \frac{n}{2} pq \end{array} \right\}$$

The mean and variance of the number of successes in all the  $n$  events are given by  $E(X_1 + X_2) = E(X_1) + E(X_2) = \frac{n}{2} p + \frac{n}{2} q = \frac{n}{2}$

$$\text{and } V(X_1 + X_2) = V(X_1) + V(X_2) = \frac{n}{2} pq + \frac{n}{2} qp = npq,$$

since the first and second half of events are independent.

Hence the variance is the same as if the probability of success in all the  $n$  events is  $p$ .

### EXERCISE 12(a)

1. (a) There are 2 populations and  $P_1$  and  $P_2$  are the proportion of members in the two populations belonging to 'low-income' group. It is desired to test the hypothesis  $H_0 : P_1 = P_2$ . Explain clearly, the procedure that you would follow to carry out the above test at 5% level of significance.

State the theorem on which the above test is based.

In respect of the above 2 populations, if it is claimed that  $P_1$ , the proportion of 'low-income' group in the first population is greater than  $P_2$ , how will you modify the procedure to test this claim (at 5% level) ?

(b) Take a concrete illustration and in relation to this illustration, explain the following terms :—

(i) Null hypothesis and alternative hypothesis.

(ii) Type I and Type II errors.

(iii) Critical Region.

(c) Suggest a possible source of bias in the following :

(i) The mean income per family in a certain town is sought to be estimated by sampling from motor owners.

(ii) Readers of newspapers are sampled by printing in it an invitation to them to send up their observations on some typical event.

(iii) A barrel of apples is sampled by taking a handful from the top.

(iv) A set of digits is taken by opening a telephone directory at random and choosing the telephone numbers in the order in which they appear on the page.

2. (a) Explain clearly the terms "Standard Error" and "Sampling Distribution." Show that in a series of  $n$  independent trials with constant probability  $p$  of success, the standard error of the proportion of successes is  $\sqrt{pq/n}$ , where  $q = 1 - p$ .

(b)  $n$  individuals fall into one or the other two categories with probabilities  $p$  and  $q (=1-p)$ , the number in the two categories are  $x_1$  and  $x_2$  ( $x_1 + x_2 = n$ ). Show that covariance between  $x_1$  and  $x_2$  is  $-npq$ . Hence obtain the variance of the difference  $\left(\frac{x_1}{n} - \frac{x_2}{n}\right)$ , between the proportions.

(c) Explain clearly the procedure generally followed in testing of a hypothesis. Point out the difference between one-tail and two-tail tests.

(d) What do you mean by interval estimation and how would you set up the confidence limits for a parameter from a sample ? Give the formula for 95% confidence limits for mean and proportion. What modifications do you have to make if the sampling is done from finite population, (i) without replacement, (ii) with replacement ? [Calcutta Univ. B.A. (Maths Hons.), 1988]

3.  $P_1$  and  $P_2$  are the (unknown) proportions of students wearing glasses in two universities  $A$  and  $B$ . To compare  $P_1$  and  $P_2$ , samples of sizes  $n_1$  and  $n_2$  are taken from the two populations and the number of students wearing glasses is found to be  $x_1$  and  $x_2$  respectively. Suggest an unbiased estimate of  $(P_1 - P_2)$  and obtain its sampling distribution when  $n_1$  and  $n_2$  are large. Hence explain how to test the hypothesis that  $P_1 = P_2$ .

4. (a) A coin is tossed 10,000 times and it turns up head 5,195 times. Discuss whether the coin may be regarded as unbiased one, explaining briefly the theoretical principles you would use for this purpose. (Ans. No.)

(b) A biased coin was thrown 400 times and head resulted 240 times. Find the standard error of the observed proportion of heads and deduce that the probability of getting a head in a single throw of the coin lies almost certainly between 0.53 and 0.67. (Ans. 0.02445).

(c) Experience has shown that 20% of a manufactured product is of the top quality. In one day's production of 400 articles only 50 are of top quality. Show that either the production of the day taken was not a representative sample or the hypothesis of 20% was wrong. (Ans. Z = 3.75)

5. (a) In a large consignment of oranges a random sample of 64 oranges revealed that 14 oranges were bad. Is it reasonable to assume that 20% of the oranges were bad?

(b) By a mobile court checking in certain buses it was found that out of 1000 people checked on a certain day at Red Fort, 10 persons were found to be ticketless travellers. If daily 1 lakh passengers travel by the buses, find out the estimated limits to the ticketless travellers. (Ans. 997 to 1003)

(c) In a random sample of 81 items taken from a large consignment some were found to be defective. If the standard error of the proportion of defective items in the sample is  $1/18$ , find 95% confidence limits of the percentage of defective items in the consignment.

[*Madras Univ. B.Sc. (Stat. Main), 1991*]

6. (a) In some dice throwing experiments Weldon threw dice 75,145 times and of these 49,152 yielded a 4, 5 or 6. Is this consistent with the hypothesis that the dice was unbiased?

**Hint.**  $H_0$  : Dice is unbiased, i.e.,  $P = \frac{3}{6} = \frac{1}{2} = 0.5$ ;  $H_1 : P \neq \frac{1}{2}$

**Test Statistic.** Under  $H_0$ ,  $Z = \frac{p - P}{\sqrt{PQ/n}} = \frac{0.654 - 0.5}{\sqrt{0.5 \times 0.5/75145}} = \frac{0.154}{0.0018}$

**Ans. No.**

(b) 1,000 apples are taken from a large consignment and 100 are found to be bad. Estimate the percentage of bad apples in the consignment and assign the limits within which the percentage lies.

7. (a) A personnel manager claims that 80 per cent of all single women hired for secretarial job get married and quit work within two years after they are hired. Test this hypothesis at 5% level of significance if among 200 such secretaries, 112 got married within two years after they were hired and quit their jobs.

(b) A manufacturer claimed that at least 98% of the steel pipes which he supplied to a factory conformed to specifications. An examination of a sample of 500 pieces of pipes revealed that 30 were defective. Test this claim at a significance level of (i) 0.05, (ii) 0.01.

**Hint.**  $X$  = No. of pipes conforming to specifications in the sample.

$$= 500 - 30 = 470$$

$p$  = Sample proportion of pipes conforming to specifications

$$= \frac{470}{500} = 0.94$$

$H_0 : P = 0.98$ , i.e., the proportion of pipes conforming to specifications in the lot is 98%.

$H_1 : P < 0.98$  (Left-tail alternative)

**Test Statistic.**  $Z = \frac{p - P}{\sqrt{PQ/n}} = \frac{0.94 - 0.98}{\sqrt{0.98 \times 0.02/500}}$

(c) A social worker believes that fewer than 25% of the couples in a certain area ever used any form of birth control. A random sample of 120 couples was

contacted. Twenty of them said they had used some method of birth control. Comment on the social worker's belief.

$$H_0 : P = 0.25, H_1 : P < 0.25 \text{ (left-Tailed)}$$

8. In a random sample of 800 adults from the population of a certain large city, 600 are found to have dark hair. In a random sample of 1,000 adults from the habitants of another large city, 700 are dark haired. Show that the difference of the proportion of dark haired people is nearly 2.4 times the standard error of the difference for samples of above sizes.

9. (a) In a random sample of 100 men taken from village A, 60 were found to be consuming alcohol. In another sample of 200 men taken from village B, 100 were found to be consuming alcohol. Do the two villages differ significantly in respect of the proportion of men who consume alcohol ?

[*Delhi Univ. M.A. (Business Eco.), 1987*]

(b) In a random sample of 500 men from a particular district of U.P., 300 are found to be smokers. In one of 1,000 men from another district, 550 are smokers. Do the data indicate that the two districts are significantly different with respect to the prevalence of smoking among men ?

**Ans.**  $Z = 1.85$ , (not significant).

[*Delhi Univ. B.Sc., 1991*)

10. A company is considering two different television advertisements for promotion of a new product. Management believed that the advertisement A is more effective than advertisement B. Two test market areas with virtually identical consumer characteristics are selected; A is used in one area and B in other area. In a random sample of 60 customers who saw A, 18 tried the product. In another random sample of 100 customers who saw B, 22 tried the product. Does this indicate that advertisement A is more effective than advertisement B, if a 5% level of significance is used ? Given critical value at 5% level is 1.96 and at 10% level of significance is 1.645.

[*Delhi Univ. M.C.A., 1990*]

11. (a) 1,000 apples kept under one type of storage were found to show rotting to the extent of 4%. 1,500 apples kept under another kind of storage showed 3% rotting. Can it be reasonably concluded that the second type of storage is superior to the first ?

(b) In a referendum submitted to the students body at a university, 850 men and 566 women voted. 530 of the men and 304 of the women voted yes. Does this indicate a significant difference of opinion on the matter at 1% level, between men and women students. [Ans.  $Z = 3.2$ , (significant).]

(c) In a simple sample of 600 high school students from a State, 400 are found to use dot pens. In one of 900 from a neighbouring State, 450 are found to use dot pens. Do the data indicate that the States are significantly different with respect to the habit of using dot pens among the students ? (Ans. Yes.)

12. (a) A firm, manufacturing dresses for children, sent out advertisement through mail. Two groups of 1,000 each were contacted; the first group having been contacted in white covers while the second in blue covers. 20% from the first while 28% from the second replied.

Do you think that blue envelopes help the sales ?

(b) A machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?

**Hint.** We are given :  $n_1 = 500$ , and  $n_2 = 100$

$$p_1 = \frac{16}{500} = 0.032; p_2 = \frac{3}{100} = 0.030$$

**Null Hypothesis,**  $H_0 : P_1 = P_2$ , i.e., there is no significant difference in the machine before overhauling and after overhauling. In other words, the machine has not improved after overhauling.

**Alternative Hypothesis,**  $H_1 : P_2 < P_1$  or  $P_1 > P_2$ .

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{16 + 3}{500 + 100} = \frac{19}{600} = 0.032$$

$$\text{S.E. } (p_1 - p_2) = \sqrt{0.032 \times 0.968 \left( \frac{1}{500} + \frac{1}{100} \right)} = 0.0193$$

$$Z = \frac{0.032 - 0.030}{0.0193} = \frac{0.002}{0.0193} = 1.04$$

Since  $Z < 1.645$  (Right-tailed test), it is not significant at 5% level of significance.

(c) In a large city  $A$ , 25% of a random sample of 900 school boys had defective eye-sight. In another large city  $B$ , 15.5% of a random sample of 1,600 school boys had the same defect. Is this difference between the two proportions significant? (Ans. Not significant.)

13. (a) A candidate for election made a speech in city  $A$  but not in  $B$ . A sample of 500 voters from city  $A$  showed that 59.6% of the voters were in favour of him, whereas a sample of 300 voters from city  $B$  showed that 50% of the voters favoured him. Discuss whether his speech could produce any effect on voters in city  $A$ . Use 5% level.

**Ans.**  $|Z| = 2.67$ . Yes.

(b) In a large city, 16 out of a random sample of 500 men were found to be drinkers. After the heavy increase in tax on intoxicants another random sample of 100 men in the same city included 3 drinkers. Was the observed decrease in the proportion of drinkers significant after tax increase?

**Ans.**  $H_0 : P_1 = P_2$ ,  $H_1 : P_1 > P_2$ ;  $Z = 1.04$ . Not significant.

14. The sex ratio at birth is sometimes given by the ratio of male to female births instead of the proportion of male to total births. If  $z$  is the ratio,

i.e.,  $z = p/q$ , show that the standard error of  $z$  is approximately  $\frac{1}{1+z} \sqrt{\left(\frac{z}{n}\right)}$   $n$  being large, so that deviations are small compared with mean.

**12.10. Sampling of Variables.** In the present section we will discuss in detail the sampling of variables such as height, weight, age, income, etc. In the case of sampling of variables each member of the population provides the value of the variable and the aggregate of these values forms the frequency distribution of the population. From the population, a random sample of size  $n$

can be drawn by any of the sampling methods discussed before which is same as choosing  $n$  values of the given variable from the distribution.

**12-11. Unbiased Estimate for population Mean ( $\mu$ ) and Variance ( $\sigma^2$ ).** Let  $x_1, x_2, \dots, x_n$  be a random sample of size  $n$  from a large population  $X_1, X_2, \dots, X_N$  (of size  $N$ ) with mean  $\mu$  and variance  $\sigma^2$ . Then the sample mean ( $\bar{x}$ ) and variance ( $s^2$ ) are given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \text{ and } s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Now } E(\bar{x}) = E\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n E(x_i)$$

Since  $x_i$  is a sample observation from the population  $X_i$ , ( $i = 1, 2, \dots, N$ ) it can take any one of the values  $X_1, X_2, \dots, X_N$  each with equal probability  $1/N$ .

$$\begin{aligned} \therefore E(x_i) &= \frac{1}{N} X_1 + \frac{1}{N} X_2 + \dots + \frac{1}{N} X_N \\ &= \frac{1}{N} (X_1 + X_2 + \dots + X_N) = \mu \end{aligned} \quad \dots(1)$$

$$\therefore E(\bar{x}) = \frac{1}{n} \sum_{i=1}^n (\mu) = \frac{1}{n} n\mu \Rightarrow E(\bar{x}) = \mu \quad \dots(12-6)$$

Thus the sample mean ( $\bar{x}$ ) is an unbiased estimate of the population mean ( $\mu$ ).

$$\begin{aligned} \text{Now } E(s^2) &= E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right] = E\left[\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2\right] \\ &= \frac{1}{n} \sum_{i=1}^n E(x_i^2) - E(\bar{x})^2 \end{aligned} \quad \dots(2)$$

$$\begin{aligned} \text{We have } V(x_i) &= E[x_i - E(x_i)]^2 = E(x_i - \mu)^2, \quad [\text{From (1)}] \\ &= \frac{1}{N} [(X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_N - \mu)^2] = \sigma^2 \quad \dots(3) \end{aligned}$$

Also we know that

$$V(x) = E(x^2) - [E(x)]^2 \Rightarrow E(x^2) = V(x) + [E(x)]^2 \quad \dots(4)$$

In particular

$$E(x_i^2) = V(x_i) + [E(x_i)]^2 = \sigma^2 + \mu^2 \quad \dots(5)$$

Also from (4),  $E(\bar{x}^2) = V(\bar{x}) + [E(\bar{x})]^2$

But  $V(\bar{x}) = \frac{\sigma^2}{n}$ , where  $\sigma^2$  is the population variance. [c.f. § 12-13]

$$\therefore E(\bar{x}^2) = \frac{\sigma^2}{n} + \mu^2 \quad [\text{Using (12-6)}] \quad \dots(5a)$$

Substituting from (5) and (5a) in (2) we get

$$\begin{aligned}
 E(s^2) &= \frac{1}{n} \sum_{i=1}^n (\sigma^2 + \mu^2) - \left( \frac{\sigma^2}{n} + \mu^2 \right) \\
 &= \frac{1}{n} n (\sigma^2 + \mu^2) - \left( \frac{\sigma^2}{n} + \mu^2 \right) = \left( 1 - \frac{1}{n} \right) \sigma^2 \\
 &= \frac{n-1}{n} \sigma^2
 \end{aligned} \quad \dots(12.7)$$

Since  $E(s^2) \neq \sigma^2$ , sample variance is not an unbiased estimate of population variance.

From (12.7), we get

$$\begin{aligned}
 \frac{n}{n-1} E(s^2) &= \sigma^2 \Rightarrow E \left( \frac{ns^2}{n-1} \right) = \sigma^2 \\
 \Rightarrow E \left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right] &= \sigma^2 \text{ i.e., } E(S^2) = \sigma^2
 \end{aligned} \quad \dots(12.8)$$

$$\text{where } S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \dots(12.8a)$$

$\therefore S^2$  is an unbiased estimate of the population variance  $\sigma^2$ .

Aliter for  $E(s^2)$ .

$$\begin{aligned}
 s^2 &= \frac{1}{n} \left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right] = \frac{1}{n} \left[ \sum_{i=1}^n \{ (x_i - \mu) - (\bar{x} - \mu) \}^2 \right] \\
 &= \frac{1}{n} \left[ \sum_{i=1}^n (x_i - \mu)^2 + n(\bar{x} - \mu)^2 - 2(\bar{x} - \mu) \sum_{i=1}^n (x_i - \mu) \right]
 \end{aligned}$$

$$\text{But } \sum_i (x_i - \mu) = \sum_i x_i - n\mu = n\bar{x} - n\mu = n(\bar{x} - \mu)$$

$$\therefore s^2 = \frac{1}{n} \left\{ \sum_{i=1}^n (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \right\} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 - (\bar{x} - \mu)^2$$

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

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

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

**Remarks 1.** Here we see that although sample mean is an unbiased estimate of population mean, sample variance is not an unbiased estimate of population variance. However, an unbiased estimate of  $\sigma^2$  is given by  $S^2$ , given in equation (12.8a).

$S^2$  plays a very important role in sampling theory, particularly in small sampling theory. Whenever  $\sigma^2$  is not known, its estimate  $S^2$  given by (12.8a) is used for practical purposes.

$$\text{2. We have } s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\Rightarrow ns^2 = (n-1)S^2 \quad \therefore s^2 = \left(1 - \frac{1}{n}\right) S^2$$

Hence for large samples i.e., for  $n \rightarrow \infty$ , we have  $s^2 \rightarrow S^2$ . In other words, for large samples (i.e.,  $n \rightarrow \infty$ ), we may take

$$\hat{\sigma}^2 = s^2 \quad \dots (12.8b)$$

**12.12. Standard Error of Sample Mean.** The variance of the sample mean is  $\sigma^2/n$ , where  $\sigma$  is the population standard deviation and  $n$  is the size of the random sample.

The S.E. of mean of a random sample of size  $n$  from a population with variance  $\sigma^2$  is  $\sigma/\sqrt{n}$ .

**Proof.** Let  $x_i$ , ( $i = 1, 2, \dots, n$ ) be a random sample of size  $n$  from a population with variance  $\sigma^2$ , then the sample mean  $\bar{x}$  is given by

$$\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

$$\therefore V(\bar{x}) = V\left[\frac{1}{n} (x_1 + x_2 + \dots + x_n)\right] = \frac{1}{n^2} V(x_1 + x_2 + \dots + x_n)$$

$$= \frac{1}{n^2} \left[ V(x_1) + V(x_2) + \dots + V(x_n) \right],$$

the covariance terms vanish since the sample observations are independent, [c.f. Remark (ii) § 6.6]

But  $V(x_i) = \sigma^2$ , ( $i = 1, 2, \dots, n$ ) [From (3) of § 12.11]

$$\therefore V(\bar{x}) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}$$

$$\Rightarrow \text{S.E.}(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \quad \dots (12.9)$$

**12.13. Test of Significance for Single Mean.** We have proved that if  $x_i$ , ( $i = 1, 2, \dots, n$ ) is a random sample of size  $n$  from a normal population with mean  $\mu$  and variance  $\sigma^2$ , then the sample mean is distributed normally with mean  $\mu$  and variance  $\sigma^2/n$ , i.e.,  $\bar{x} \sim N(\mu, \sigma^2/n)$ . However, this result holds, i.e.,  $\bar{x} \sim N(\mu, \sigma^2/n)$ , even in random sampling from non-normal population provided the sample size  $n$  is large [c.f. Central Limit Theorem, § 8.10].

Thus for large samples, the standard normal variate corresponding to  $\bar{x}$  is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Under the *null hypothesis*,  $H_0$  that the sample has been drawn from a population with mean  $\mu$  and variance  $\sigma^2$ , i.e., there is no significant difference between the sample mean ( $\bar{x}$ ) and population mean ( $\mu$ ), the test statistic (for large samples), is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \dots(12.9a)$$

**Remarks 1.** If the population s.d.  $\sigma$  is unknown then we use its estimate provided by the sample variance given by [See (12.8b)]:

$$\hat{\sigma}^2 = s^2 \Rightarrow \hat{\sigma} = s \text{ (for large samples).}$$

**2. Confidence limits for  $\mu$ .** 95% confidence interval for  $\mu$  is given by :

$$|Z| \leq 1.96, \text{ i.e., } \left| \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right| \leq 1.96$$

$$\Rightarrow \bar{x} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1.96\sigma/\sqrt{n} \quad \dots(12.10)$$

and  $\bar{x} \pm 1.96\sigma/\sqrt{n}$  are known as 95% confidence limits for  $\mu$ . Similarly, 99% confidence limits for  $\mu$  are  $\bar{x} \pm 2.58\sigma/\sqrt{n}$  and 98% confidence limits for  $\mu$  are  $\bar{x} \pm 2.33\sigma/\sqrt{n}$ .

However, in sampling from a finite population of size  $N$ , the corresponding 95% and 99% confidence limits for  $\mu$  are respectively

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \text{ and } \bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad \dots(12.10a)$$

**3.** The confidence limits for any parameter ( $P$ ,  $\mu$ , etc.) are also known as its *fiducial limits*.

**Example 12.15.** A sample of 900 members has a mean 3.4 cms., and s.d. 2.61 cms. Is the sample from a large population of mean 3.25 cms. and s.d. 2.61 cms.?

If the population is normal and its mean is unknown, find the 95% and 98% fiducial limits of true mean.

**Solution.** *Null hypothesis*, ( $H_0$ ) : The sample has been drawn from the population with mean  $\mu = 3.25$  cms., and S.D.  $\sigma = 2.61$  cms.

*Alternative Hypothesis*,  $H_1 : \mu \neq 3.25$  (Two-tailed).

*Test Statistic.* Under  $H_0$ , the test statistic is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ (since } n \text{ is large)}$$

Here, we are given

$$\bar{x} = 3.4 \text{ cms.}, n = 900 \text{ cms.}, \mu = 3.25 \text{ cms. and } \sigma = 2.61 \text{ cms.}$$

$$Z = \frac{3.40 - 3.25}{2.61/\sqrt{900}} = \frac{0.15 \times 30}{2.61} = 1.73$$

Since  $|Z| < 1.96$ , we conclude that the data don't provide us any evidence against the null hypothesis ( $H_0$ ) which may, therefore, be accepted at 5% level of significance.

95% fiducial limits for the population mean  $\mu$  are :

$$\bar{x} \pm 1.96 \sigma / \sqrt{n} \Rightarrow 3.40 \pm 1.96 \times 2.61 / \sqrt{900}$$

$$\Rightarrow 3.40 \pm 0.1705, \text{ i.e., } 3.5705 \text{ and } 3.2295$$

98% fiducial limits for  $\mu$  are given by :

$$\bar{x} \pm 2.33 \frac{\sigma}{\sqrt{n}}, \text{ i.e., } 3.40 \pm 2.33 \times \frac{2.61}{30}$$

$$\Rightarrow 3.40 \pm 0.2027 \text{ i.e., } 3.6027 \text{ and } 3.1973$$

**Remark.** 2.33 is the value  $z_1$  of  $Z$  from standard normal probability integrals, such that  $P(|Z| > z_1) = 0.98 \Rightarrow P(Z > z_1) = 0.49$ .

**Example 12-16.** An insurance agent has claimed that the average age of policyholders who insure through him is less than the average for all agents, which is 30.5 years.

A random sample of 100 policyholders who had insured through him gave the following age distribution :

Age last birthday	No. of persons
16–20	12
21–25	22
26–30	20
31–35	30
36–40	16

Calculate the arithmetic mean and standard deviation of this distribution and use these values to test his claim at the 5% level of significance. You are given that  $Z(1.645) = 0.95$ .

**Solution.** Null Hypothesis,  $H_0 : \mu = 30.5$  years, i.e., the sample mean ( $\bar{x}$ ) and population mean ( $\mu$ ) do not differ significantly.

Alternative Hypothesis,  $H_1 : \mu < 30.5$  years (Left-tailed alternative).

#### CALCULATIONS FOR SAMPLE MEAN AND S.D.

Age last birthday	No. of persons ( $f$ )	Mid-point $x$	$d = \frac{x - 28}{5}$	$fd$	$fd^2$
16–20	12	18	-2	-24	48
21–25	22	23	-1	-22	22
26–30	20	28	0	0	0
31–35	30	33	1	30	30
36–40	16	38	2	32	64
Total	$N = 100$			$\sum fd = 16$	$\sum fd^2 = 164$

$$\bar{x} = 28 + \frac{5 \times 16}{100} = 28.8 \text{ years} \quad s = 5 \times \sqrt{\frac{164}{100} - \left(\frac{16}{100}\right)^2} = 6.35 \text{ years}$$

Since the sample is large,  $\hat{\sigma} \approx s = 6.35$  years.

*Test Statistic.* Under  $H_0$ , the test statistic is

$$Z = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim N(0, 1), \text{ (since sample is large).}$$

$$\text{Now } Z = \frac{28.8 - 30.5}{6.35/\sqrt{100}} = \frac{-1.7}{0.635} = -2.681$$

*Conclusion.* Since computed value of  $Z = -2.681 < -1.645$  or  $|Z| = 2.681 > 1.645$ , it is significant at 5% level of significance. Hence we reject the null hypothesis  $H_0$  (Accept  $H_1$ ) at 5% level of significance and conclude that the insurance agent's claim that the average age of policyholders who insure through him is less than the average for all agents, is valid.

**Example 12.17.** As an application of Central Limit Theorem, show that if  $E$  is such that  $P(|\bar{X} - \mu| < E) > 0.95$ , then the minimum sample size  $n$  is given by  $n = \frac{(1.96)^2 \sigma^2}{E^2}$ , where  $\mu$  and  $\sigma^2$  are the mean and variance respectively of the population and  $\bar{X}$  is the mean of the random sample.

**Solution.** By Central Limit Theorem, we know that  $\bar{X} \sim N(\mu, \sigma^2/n)$  asymptotically i.e., for large  $n$ .

$$\therefore Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ asymptotically i.e., for large } n.$$

From normal probability tables, we have

$$P(|Z| \leq 1.96) = 0.95$$

$$\Rightarrow P\left[\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq 1.96\right] = 0.95$$

$$\Rightarrow P\left[|\bar{X} - \mu| \leq 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95 \quad \dots (*)$$

We are given that

$$P(|\bar{X} - \mu| < E) > 0.95 \quad \dots (**)$$

From (\*) and (\*\*), we have

$$E > \frac{1.96\sigma}{\sqrt{n}} \Rightarrow n > \frac{(1.96)^2 \sigma^2}{E^2} = \frac{3.84\sigma^2}{E^2}$$

Hence minimum sample size  $n$  for estimating  $\mu$  with 95% confidence coefficient is given by  $n = 3.84 \sigma^2/E^2$ , where  $E$  is the permissible error.

**Remark.** The minimum sample size for estimating  $\mu$  with confidence coefficient  $(1 - \alpha)$  is given by  $\sigma^2 z_{\alpha/2}^2/E^2$ , where  $z_{\alpha/2}$  is the significant value of  $Z$  at level of significance  $\alpha$  and  $E$  is the permissible error in the estimate.

Arguing similarly, the minimum sample size for estimating population proportion  $P$  with confidence coefficient  $(1 - \alpha)$  is given by  $n = PQ z_{\alpha}^2/E^2$ , where  $z_{\alpha}$  is the significant value of Z at ' $\alpha$ ' level of significance and  $E$  is the permissible error in the estimate. If  $P$  is unknown, we may use  $\hat{P} = p$ .

**Example 12-18.** The mean muscular endurance score of a random sample of 60 subjects was found to be 145 with a s.d. of 40. Construct a 95% confidence interval for the true mean. Assume the sample size to be large enough for normal approximation. What size of sample is required to estimate the mean within 5 of the true mean with a 95% confidence?

[Calicut Univ. B.Sc. (Main Stat.) 1989]

**Solution.** We are given :  $n = 60$ ,  $\bar{x} = 145$  and  $s = 40$ .

95% confidence limits for true mean ( $\mu$ ) are :

$$\begin{aligned}\bar{x} &\pm 1.96 s/\sqrt{n} & (\sigma^2 = s^2, \text{ since sample is large}) \\ &= 145 \pm \frac{1.96 \times 40}{\sqrt{60}} = 145 \pm \frac{78.4}{7.75} = 145 \pm 10.12 = 134.88, 155.12\end{aligned}$$

Hence 95% confidence interval for  $\mu$  is (134.88, 155.12). In the notations of Example 12-17, we have

$$\begin{aligned}n &= \left( \frac{z_{0.05} \cdot \sigma}{E} \right)^2 = \left( \frac{1.96 \times 40}{5} \right)^2 \\ &[ \because z_{0.05} = 1.96, \sigma = s = 40 \text{ and } |\bar{x} - \mu| < 5 = E ] \\ &= (15.68)^2 = 245.86 \approx 246.\end{aligned}$$

**Example 12-19.** The standard deviation of a population is 2.70 inches. Find the probability that in a random sample of size 66 (i) the sample mean will differ from the population mean by 0.75 inch or more and (ii) the sample mean will exceed the population mean by 0.75 inch or more (given that the value of the standard normal probability integral from 0 to 2.25 is 0.4877).

**Solution.** Here we are given  $n = 66$ ,  $\sigma = 2.70$  inches. Since  $n$  is large, the sample mean  $\bar{x} \sim N(\mu, \sigma^2/n)$ .

$$\therefore Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \dots (*)$$

We want

$$\begin{aligned}(i) \quad P[|\bar{x} - \mu| \geq 0.75] &= 1 - P[|\bar{x} - \mu| < 0.75] \\ &= 1 - P\left[\left|\frac{\sigma}{\sqrt{n}}Z\right| < 0.75\right] & [\text{From } (*)] \\ &= 1 - P\left[|Z| < 0.75 \frac{\sqrt{n}}{\sigma}\right] \\ &= 1 - 2P\left[0 < Z < 0.75 \frac{\sqrt{n}}{\sigma}\right]\end{aligned}$$

$$\begin{aligned}
 &= 1 - 2 P \left[ 0 < Z < 0.75 \times \frac{\sqrt{66}}{2.70} \right] \\
 &= 1 - 2 P \left[ 0 < Z < \frac{0.75 \times 8.124}{2.70} \right] \\
 &= 1 - 2 P[0 < Z < 2.25] = 1 - 2 \times 0.4877 = 0.0246 \\
 (ii) P(\bar{x} - \mu > 0.75) &= P(Z > 0.75 \sqrt{n}/\sigma) = P(Z > 2.25) \\
 &= 0.5 - P(0 < Z < 2.25) = 0.5 - 0.4877 = 0.0123
 \end{aligned}$$

**Example 12-20.** A normal population has a mean of 0.1 and standard deviation of 2.1. Find the probability that mean of a sample of size 900 will be negative. [Delhi Univ. B.Sc. (Stat. Hons.), 1986]

**Solution.** Here we are given that  $X \sim N(\mu, \sigma^2)$ , where  $\mu = 0.1$  and  $\sigma = 2.1$  and  $n = 900$ .

Since  $X \sim N(\mu, \sigma^2)$ , the sample mean  $\bar{x} \sim N(\mu, \sigma^2/n)$ . The standard normal variate corresponding to  $\bar{x}$  is given by :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0.1}{2.1/30} = \frac{\bar{x} - 0.1}{0.07}$$

$$\therefore \bar{x} = 0.1 + 0.07Z, \text{ where } Z \sim N(0, 1)$$

The required probability  $p$ , that the sample mean is negative is given by :

$$\begin{aligned}
 p &= P(\bar{x} < 0) = P(0.1 + 0.07Z < 0) \\
 &= P \left( Z < \frac{-0.10}{0.07} \right) = P(Z < -1.43) = P(Z \geq 1.43) \\
 &= 0.5 - P(0 < Z < 1.43) = 0.5 - 0.4236 = 0.0764
 \end{aligned}$$

(From Normal Probability Tables).

**Example 12-21.** The guaranteed average life of a certain type of electric light bulbs is 1000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90 per cent of the bulbs do not fall short of the guaranteed average by more than 2.5 per cent. What must be the minimum size of the sample ? [Madras Univ. B.Sc., Oct. 1991]

**Solution.** Here  $\mu = 1000$  hours,  $\sigma = 125$  hours.

Since we do not want the sample mean to be less than the guaranteed average mean ( $\mu = 1000$ ) by more than 2.5%, we should have

$$\bar{x} > 1000 - 2.5\% \text{ of } 1000 \Rightarrow \bar{x} > 1000 - 25 = 975$$

Let  $n$  be the given sample size. Then

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ since sample is large.}$$

$$\text{We want } Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \frac{975 - 1000}{125/\sqrt{n}} > -\frac{\sqrt{n}}{5} \quad (\because \bar{x} > 975)$$

According to the given condition, we have

$$\begin{aligned} P(Z > -\sqrt{n/5}) &= 0.90 \Rightarrow P(0 < Z < \sqrt{n/5}) = 0.40 \\ \therefore \sqrt{n/5} &= 1.28 \quad (\text{From Normal Probability Tables}) \\ \Rightarrow n &= 25 \times (1.28)^2 = 41 \text{ (approx)} \end{aligned}$$

**Example 12-22.** A survey is proposed to be conducted to know the annual earnings of the old Statistics graduates of Delhi University. How large should the sample be taken in order to estimate the mean annual earnings within plus and minus Rs. 1,000 at 95% confidence level? The standard deviation of the annual earnings of the entire population is known to be Rs. 3,000.

**Solution.** We are given :  $\sigma = \text{Rs. } 3,000$ .

$$\text{We want : } P[|\bar{x} - \mu| < 1,000] = 0.95 \quad \dots(*)$$

We know that, in sampling from normal population or for large samples from any population  $\bar{X} \sim N(\mu, \sigma^2/n)$ . Hence from normal probability tables, we have :

$$\begin{aligned} P[|Z| \leq 1.96] &= 0.95 \\ \Rightarrow P\left[\left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \leq 1.96\right] &= 0.95 \\ \Rightarrow P[|\bar{x} - \mu| \leq 1.96 \times (\sigma/\sqrt{n})] &= 0.95 \quad \dots(**) \end{aligned}$$

From (\*) and (\*\*), we get

$$\begin{aligned} \frac{1.96 \times \sigma}{\sqrt{n}} &= 1000 \Rightarrow \frac{1.96 \times 3000}{\sqrt{n}} = 1000 \\ \therefore n &= (1.96 \times 3)^2 = (5.88)^2 = 34.56 \approx 35 \end{aligned}$$

**Aliter.** Using Remark to Example 12-17,

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{1.96 \times 3,000}{1,000}\right)^2 \approx 35.$$

**12-14. Test of Significance for Difference of Means.** Let  $\bar{x}_1$  be the mean of a random sample of size  $n_1$  from a population with mean  $\mu_1$  and variance  $\sigma_1^2$  and let  $\bar{x}_2$  be the mean of an independent random sample of size  $n_2$  from another population with mean  $\mu_2$  and variance  $\sigma_2^2$ . Then, since sample sizes are large,

$$\bar{x}_1 \sim N(\mu_1, \sigma_1^2/n_1) \text{ and } \bar{x}_2 \sim N(\mu_2, \sigma_2^2/n_2)$$

Also  $\bar{x}_1 - \bar{x}_2$ , being the difference of two independent normal variates is also a normal variate. The Z (S.N.V.) corresponding to  $\bar{x}_1 - \bar{x}_2$  is given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.(\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Under the null hypothesis  $H_0 : \mu_1 = \mu_2$ , i.e., there is no significant difference between the sample means, we get

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2 = 0;$$

$$V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

the covariance term vanishes, since the sample means  $\bar{x}_1$  and  $\bar{x}_2$  are independent.

Thus under  $H_0 : \mu_1 = \mu_2$ , the test statistic becomes (for large samples),

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \sim N(0, 1) \quad \dots(12.11)$$

**Remarks 1.** If  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , i.e., if the samples have been drawn from the populations with common S.D.  $\sigma$ , then under  $H_0 : \mu_1 = \mu_2$ ,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{(1/n_1 + 1/n_2)}} \sim N(0, 1) \quad \dots[12.11(a)]$$

2. If in (12.11a),  $\sigma$  is not known, then its estimate based on the sample variances is used. If the sample sizes are not sufficiently large, then an unbiased estimate of  $\sigma^2$  is given by

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)},$$

since

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)E(S_1^2) + (n_2 - 1)E(S_2^2)] \\ &= \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2] = \sigma^2 \end{aligned}$$

But since sample sizes are large,  $S_1^2 \approx s_1^2$ ,  $S_2^2 \approx s_2^2$ ,  $n_1 - 1 \approx n_1$ ,  $n_2 - 1 \approx n_2$ . Therefore in practice, for large samples, the following estimate of  $\sigma^2$  without any serious error is used :

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} \quad \dots[12.11(b)]$$

However, if sample sizes are small, then a small sample test,  $t$ -test for difference of means (c.f. Chapter 14) is to be used.

3. If  $\sigma_1^2 \neq \sigma_2^2$  and  $\sigma_1$  and  $\sigma_2$  are not known, then they are estimated from sample values. This results in some error, which is practically immaterial, if samples are large. These estimates for large samples are given by

$$\begin{cases} \hat{\sigma}_1^2 = S_1^2 \approx s_1^2 \\ \hat{\sigma}_2^2 = S_2^2 \approx s_2^2 \end{cases} \quad (\text{since samples are large}).$$

In this case, (12.11) gives

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \sim N(0, 1) \quad \dots[12.11(c)]$$

**Example 12.23.** The means of two single large samples of 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2.5 inches ? (Test at 5% level of significance).

**Solution.** We are given :

$$n_1 = 1000, n_2 = 2000; \bar{x}_1 = 67.5 \text{ inches}, \bar{x}_2 = 68.0 \text{ inches}.$$

**Null hypothesis,**  $H_0 : \mu_1 = \mu_2$  and  $\sigma = 2.5$  inches, i.e., the samples have been drawn from the same population of standard deviation 2.5 inches.

**Alternative Hypothesis,**  $H_1 : \mu_1 \neq \mu_2$  (Two tailed.)

**Test Statistic.** Under  $H_0$ , the test statistic is (since samples are large)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1)$$

$$\text{Now } Z = \frac{67.5 - 68.0}{2.5 \times \sqrt{\frac{1}{1000} + \frac{1}{2000}}} = \frac{-0.5}{2.5 \times 0.0387} = -5.1$$

**Conclusion.** Since  $|Z| > 3$ , the value is highly significant and we reject the null hypothesis and conclude that samples are certainly not from the same population with standard deviation 2.5.

**Example 12.24.** In a survey of buying habits, 400 women shoppers are chosen at random in super market 'A' located in a certain section of the city. Their average weekly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For 400 women shoppers chosen at random in super market 'B' in another section of the city, the average weekly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test at 1% level of significance whether the average weekly food expenditure of the two populations of shoppers are equal.

**Solution.** In the usual notations, we are given that

$$n_1 = 400, \quad \bar{x}_1 = \text{Rs. } 250, \quad s_1 = \text{Rs. } 40$$

$$n_2 = 400, \quad \bar{x}_2 = \text{Rs. } 220, \quad s_2 = \text{Rs. } 55$$

**Null hypothesis,**  $H_0 : \mu_1 = \mu_2$ , i.e., the average weekly food expenditures of the two populations of shoppers are equal.

**Alternative Hypothesis,**  $H_1 : \mu_1 \neq \mu_2$ . (Two-tailed)

**Test Statistic.** Since samples are large, under  $H_0$ , the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left( \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)}} \sim N(0, 1)$$

Since  $\sigma_1$  and  $\sigma_2$ , the population standard deviations are not known, we can take for large samples (c.f. § 12.15, Remark 3) :

$$\hat{\sigma}_1^2 = s_1^2 \text{ and } \hat{\sigma}_2^2 = s_2^2$$

and then  $Z$  is given by

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}} = \frac{250 - 220}{\sqrt{\left\{ \frac{(40)^2}{400} + \frac{(55)^2}{400} \right\}}} = 8.82 \text{ (approx.)}$$

**Conclusion.** Since  $|Z|$  is much greater than 2.58, the null hypothesis ( $\mu_1 = \mu_2$ ) is rejected at 1% level of significance and we conclude that the average weekly expenditures of two populations of shoppers in markets  $A$  and  $B$  differ significantly.

**Example 12.25.** The average hourly wage of a sample of 150 workers in a plant 'A' was Rs. 2.56 with a standard deviation of Rs. 1.08. The average wage of a sample of 200 workers in plant 'B' was Rs. 2.87 with a standard deviation of Rs. 1.28. Can an applicant safely assume that the hourly wages paid by plant 'B' are higher than those paid by plant 'A'?

**Solution.** Let  $X_1$  and  $X_2$  denote the hourly wages (in Rs.) of workers in plant  $A$  and plant  $B$  respectively. Then we are given :

$$n_1 = 150, \bar{x}_1 = 2.56, s_1 = 1.08 = \hat{\sigma}_1$$

$$n_2 = 200, \bar{x}_2 = 2.87, s_2 = 1.28 = \hat{\sigma}_2$$

**Null hypothesis,**  $H_0 : \mu_1 = \mu_2$ , i.e., there is no significant difference between the mean level of wages of workers in plant  $A$  and plant  $B$ .

**Alternative hypothesis,**  $H_1 : \mu_2 > \mu_1$  i.e.,  $\mu_1 < \mu_2$  (Left-tailed test)

**Test Statistic.** Under  $H_0$ , the test statistic (for large samples) is :

$$\begin{aligned} Z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1) \\ \therefore Z &= \frac{2.56 - 2.87}{\sqrt{\left\{\frac{(1.08)^2}{150} + \frac{(1.28)^2}{200}\right\}}} = \frac{-0.31}{\sqrt{0.016}} = \frac{-0.31}{0.126} = -2.46. \end{aligned}$$

**Critical region.** For a one-tailed test, the critical value of  $Z$  at 5% level of significance is 1.645. The critical region for left-tailed test thus consists of all values of  $Z \leq -1.645$ .

**Conclusion.** Since calculated value of  $Z$  (-2.46) is less than critical value (-1.645), it is significant at 5% level of significance. Hence the null hypothesis is rejected at 5% level of significance and we conclude that the average hourly wages paid by plant 'B' are certainly higher than those paid by plant 'A'.

**Example 12.26.** In a certain factory there are two independent processes manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 ozs. with a standard deviation of 12 ozs. while the corresponding figures in a sample of 400 items from the other process are 124 and 14. Obtain the standard error of difference between the two sample means. Is this difference significant? Also find the 99% confidence limits for the difference in the average weights of items produced by the two processes respectively.

**Solution.** We have

$$\begin{aligned} n_1 &= 250, \bar{x}_1 = 120 \text{ oz.}, s_1 = 12 \text{ oz.} = \hat{\sigma}_1 \\ n_2 &= 400, \bar{x}_2 = 124 \text{ oz.}, s_2 = 14 \text{ oz.} = \hat{\sigma}_2 \end{aligned} \quad \left. \right\} \text{ (since samples are large).}$$

$$\begin{aligned} S.E. (\bar{x}_1 - \bar{x}_2) &= \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)} = \sqrt{(s_1^2/n_1) + (s_2^2/n_2)} \\ &= \sqrt{\left(\frac{144}{250} + \frac{196}{400}\right)} = \sqrt{(0.576 + 0.490)} = 1.034 \end{aligned}$$

*Null Hypothesis.*,  $H_0 : \mu_1 = \mu_2$ , i.e., the sample means do not differ significantly.

*Alternative Hypothesis.*,  $H_1 : \mu_1 \neq \mu_2$  (Two-tailed).

*Test Statistic.* Under  $H_0$ , the test statistic is :

$$\begin{aligned} Z &= \frac{\bar{x}_1 - \bar{x}_2}{S.E. (\bar{x}_1 - \bar{x}_2)} = \frac{120 - 124}{1.034} \sim N(0, 1) \\ \therefore |Z| &= \frac{4}{1.034} = 3.87 \end{aligned}$$

*Conclusion.* Since  $|Z| > 3$ , the null hypothesis is rejected and we conclude that there is significant difference between the sample means.

99% confidence limits for  $|\mu_1 - \mu_2|$ , i.e., for the difference in the average weights of items produced by two processes, are

$$\begin{aligned} |\bar{x}_1 - \bar{x}_2| \pm 2.58 \text{ S.E.} (\bar{x}_1 - \bar{x}_2) &= 4 \pm 2.58 \times 1.034 \\ &= 4 \pm 2.67 \text{ (approx.)} = 6.67 \text{ and } 1.33 \\ \therefore 1.33 < |\mu_1 - \mu_2| < 6.67 \end{aligned}$$

**Example 12.27.** The mean height of 50 male students who showed above average participation in college athletics was 68.2 inches with a standard deviation of 2.5 inches; while 50 male students who showed no interest in such participation had a mean height of 67.5 inches with a standard deviation of 2.8 inches.

(i) Test the hypothesis that male students who participate in college athletics are taller than other male students.

(ii) By how much should the sample size of each of the two groups be increased in order that the observed difference of 0.7 inches in the mean heights be significant at the 5% level of significance.

**Solution.** Let  $X_1$  and  $X_2$  denote the height (in inches) of athletic participants and non-athletic participants respectively. In the usual notations, we are given :

$$n_1 = 50, \bar{x}_1 = 68.2, s_1 = 2.5; n_2 = 50, \bar{x}_2 = 67.5, s_2 = 2.8$$

*Null hypothesis.*,  $H_0 : \mu_1 = \mu_2$ .

*Alternative hypothesis.*,  $H_1 : \mu_1 > \mu_2$  (Right-tailed).

*Test Statistic.* Under  $H_0$ , the test statistic for large samples is :

$$\begin{aligned} Z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1) \\ \therefore Z &= \frac{68.2 - 67.5}{\sqrt{\left\{\frac{(2.5)^2}{50} + \frac{(2.8)^2}{50}\right\}}} = \frac{0.7}{\sqrt{0.282}} = \frac{0.7}{0.53} = 1.32 \end{aligned}$$

For a right-tailed test, the critical (significant) value of  $Z$  at 5% level of significance is 1.645.

(i) Since the calculated value of  $Z(1.32)$  is less than the critical value (1.645), it is not significant at 5% level of significance. Hence the null hypothesis is accepted and we conclude that the college athletes are not taller than other male students.

(ii) The difference between the mean heights of two groups, each of size  $n$  will be significant at 5% level of significance if  $Z \geq 1.645$

$$\begin{aligned} &\Rightarrow \frac{68.2 - 67.5}{\sqrt{\left[ \frac{(2.5)^2}{n} + \frac{(2.8)^2}{n} \right]}} \geq 1.645 \\ &\Rightarrow \frac{0.7}{\sqrt{14.09/n}} \geq 1.645 \Rightarrow \frac{0.7}{3.754/\sqrt{n}} \geq 1.645 \\ &\Rightarrow n \geq \left( \frac{1.645 \times 3.754}{0.7} \right)^2 = (8.8219)^2 = 77.83 \approx 78 \end{aligned}$$

Hence the sample size of each of the two groups should be increased by at least  $78 - 50 = 28$ , in order that the difference between the mean heights of the two groups is significant.

**12-15. Test of Significance for the Difference of Standard Deviations.** If  $s_1$  and  $s_2$  are the standard deviations of two independent samples, then under null hypothesis,  $H_0: \sigma_1 = \sigma_2$ , i.e., i.e., the sample standard deviations don't differ significantly, the statistic

$$Z = \frac{s_1 - s_2}{S.E.(s_1 - s_2)} \sim N(0, 1) \text{ for large samples.}$$

But in case of large samples, the S.E. of the difference of the sample standard deviations is given by

$$\begin{aligned} S.E.(s_1 - s_2) &= \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}} \\ \therefore Z &= \frac{s_1 - s_2}{\sqrt{\left( \frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2} \right)}} \sim N(0, 1) \quad \dots(12-12) \end{aligned}$$

$\sigma_1^2$  and  $\sigma_2^2$  are usually unknown and for large samples, we use their estimates given by the corresponding sample variances. Hence the test statistic reduces to

$$Z = \frac{s_1 - s_2}{\sqrt{\left( \frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2} \right)}} \sim N(0, 1) \quad \dots(12-13)$$

**Example 12-18.** Random samples drawn from two countries gave the following data relating to the heights of adult males :

	<i>Country A</i>	<i>Country B</i>
<i>Mean height (in inches)</i>	67.42	67.25
<i>Standard deviation (in inches)</i>	2.58	2.50
<i>Number in samples</i>	1000	1200

(i) Is the difference between the means significant?

(ii) Is the difference between the standard deviations significant?

**Solution.** We are given :

$$n_1 = 1000, \quad \bar{x}_1 = 67.42 \text{ inches}, \quad s_1 = 2.58 \text{ inches},$$

$$n_2 = 1200, \quad \bar{x}_2 = 67.25 \text{ inches}, \quad s_2 = 2.50 \text{ inches}.$$

As in the last examples (since sample sizes are large), we can take

$$\hat{\sigma}_1 = s_1 = 2.58, \quad \hat{\sigma}_2 = s_2 = 2.50$$

(i)  $H_0 : \mu_1 = \mu_2$ , i.e., the sample means do not differ significantly.

$$H_1 : \mu_1 \neq \mu_2 \text{ (Two tailed).}$$

Under the Null hypothesis  $H_0$ , the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \sim N(0, 1), \text{ since samples are large.}$$

$$\text{Now } Z = \frac{67.42 - 67.25}{\sqrt{\left\{ \frac{(2.58)^2}{1000} + \frac{(2.50)^2}{1200} \right\}}} = \frac{0.17}{\sqrt{\left( \frac{6.66}{1000} + \frac{6.25}{1200} \right)}} = 1.56$$

**Conclusion.** Since  $|Z| < 1.96$ , null hypothesis may be accepted at 5% level of significance and we may conclude that there is no significant difference between the sample means.

(ii) Under  $H_0$  : that there is no significant difference between sample standard deviations,

$$Z = \frac{s_1 - s_2}{S.E. (s_1 - s_2)} \sim N(0, 1), \text{ since samples are large.}$$

$$\text{Now } S.E. (s_1 - s_2) = \sqrt{\left( \frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2} \right)} = \sqrt{\left( \frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2} \right)},$$

if  $\sigma_1$  and  $\sigma_2$  are not known and  $\hat{\sigma}_1 = s_1$ ,  $\hat{\sigma}_2 = s_2$ .

$$\therefore S.E. (s_1 - s_2) = \sqrt{\left\{ \frac{(2.58)^2}{2 \times 1000} + \frac{(2.50)^2}{2 \times 1200} \right\}} = 0.07746$$

$$\text{Hence } Z = \frac{2.58 - 2.50}{0.07746} = \frac{0.08}{0.07746} = 1.03$$

**Conclusion.** Since  $|Z| < 1.96$ , the data don't provide us any evidence against the null hypothesis which may be accepted at 5% level of significance. Hence the sample standard deviations do not differ significantly.

**Example 12-29.** Two populations have their means equal, but S.D. of one is twice the other. Show that in the samples of size 2000 from each drawn under simple sampling conditions, the difference of means will, in all probability, not exceed  $0.15\sigma$ , where  $\sigma$  is the smaller S.D. What is the probability that the difference will exceed half this amount?

**Solution.** Let the standard deviations of the two populations be  $\sigma$  and  $2\sigma$  respectively and let  $\mu$  be the mean of each of the two populations. Also we are given  $n_1 = n_2 = 2000$ . If  $\bar{x}_1$  and  $\bar{x}_2$  be the two sample means then, since samples are large,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E. (\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Now  $E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu - \mu = 0$  and

$$S.E. (\bar{x}_1 - \bar{x}_2) = \sqrt{\left\{ \frac{\sigma^2}{n_1} + \frac{(2\sigma)^2}{n_2} \right\}} = \sigma \cdot \sqrt{\left( \frac{1}{2000} + \frac{4}{2000} \right)} = 0.05\sigma$$

$$\therefore Z = \frac{\bar{x}_1 - \bar{x}_2}{S.E. (\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Under simple sampling conditions, we should in all probability have

$$|Z| < 3 \Rightarrow |\bar{x}_1 - \bar{x}_2| < 3 S.E. (\bar{x}_1 - \bar{x}_2)$$

$$\Rightarrow |\bar{x}_1 - \bar{x}_2| < 0.15\sigma,$$

which is the required result.

We want  $p = P\left[|\bar{x}_1 - \bar{x}_2| > \frac{1}{2} \times 0.15\sigma\right]$

$$\begin{aligned} \therefore p &= P[0.05\sigma |Z| > 0.075\sigma] \quad \left[ \because Z = \frac{\bar{x}_1 - \bar{x}_2}{0.05\sigma} \sim N(0, 1) \right] \\ &= P[|Z| > 1.5] = 1 - P[|Z| \leq 1.5] \\ &= 1 - 2P(0 \leq Z \leq 1.5) = 1 - 2 \times 0.4332 = 0.1336 \end{aligned}$$

## EXERCISE 12-2

- Define sampling distribution and standard error. Obtain standard error of mean when population is large.
- Find the standard error of a linear function of a number of variables. Deduce the standard error of the mean of  $n$  uncorrelated variables following the same distribution.
- Derive the expressions for the standard error of
  - the mean of a random sample of size  $n$ , and
  - the difference of the means of two independent random samples of sizes  $n_1$  and  $n_2$ .
- (a) What is meant by a statistical hypothesis? What are the two types of errors of decision that arise in testing a hypothesis? Briefly explain how a statistical hypothesis is tested.

The manufacturer of television tubes knows from past experience that the

average life of a tube is 2,000 hours with a standard deviation of 200 hours. A sample of 100 tubes has an average life of 1950 hours. Test at the 0.05 level of significance if this sample came from a normal population of mean 2,000 hours.

State your null and alternative hypothesis and indicate clearly whether a one-tail or a two-tail test is used and why ? Is the result of the test significant ?

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1990*]

(b) A sample of 100 items, drawn from a universe with mean value 64 and S.D. 3 has a mean value 63.5. Is the difference in the means significant ? What will be your inference, if the sample had 200 items ?

[*Madras Univ. B.E., Nov. 1990*]

(c) A sample of 400 individuals is found to have a mean height of 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height of 67.39 inches and standard deviation 1.30 inches ?

Ans. Yes,  $Z = 1.23$ .

(d) The mean breaking strength of cables supplied by a manufacturer is 1800 with a standard deviation 100. By a new technique in the manufacturing process it is claimed that the breaking strength of the cables has increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 0.01 level of significance ?

Ans.  $H_0 : \mu = 1800, H_1 : \mu > 1800, Z = 3.535$ .

(e) An ambulance service claims that it takes on the average 8.9 minutes to reach its destination in emergency calls. To check on this claim, the agency which licenses ambulance services has them timed on 50 emergency calls, getting a mean of 9.3 minutes with a standard deviation of 1.6 minutes. What can they conclude at the level of significance  $\alpha = 0.05$  ?

Ans.  $Z = 1.768$ .

(f) A paper mill in U.P. has agreed to buy waste paper for recycling from a waste collection firm, under the agreement that the waste collection firm will supply the waste paper in packages of 300 kg each, for which the paper mill will pay by the package. To speed up their work the waste collection firm is making packages by some *approximation* procedure. The paper mill does not object to this procedure as long as it gets 300 kg. per package on the average. The waste collection firm has an interest not to exceed 300 kg. per package, because it is not being paid for more, and not to go under 300 kg. because the paper mill might terminate the agreement if it does. To estimate the mean weight of waste paper per package, the waste collection firm weighed 75 randomly selected packages and found that the mean weight was 290 kg and standard deviation was 15 kg. Can we infer that the mean weight per package in the entire supply was 300 kg ?

[*Delhi Univ. M.A. (Eco.), 1987*]

Ans.  $H_0 : \mu = 300 \text{ kg}; H_1 : \mu \neq 300 \text{ kg. (Two-tailed)}$ .

$$Z = \frac{290 - 300}{15/\sqrt{75}} = 5.77; \text{ Significant.}$$

(g) The wages of a factory's workers are assumed to be normally distributed with mean  $\mu$  and variance 25. A random sample of 25 workers gives the total wages equal to 1250 units.

Test the hypothesis :  $\mu = 52$ , against the alternative :  $\mu = 49$ , at 1% level of significance.

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2.32} \exp\left(-\frac{1}{2}\mu^2\right) d\mu = 0.01.$$

[Calcutta Univ. B.Sc.(Maths. Hons.), 1988]

**Ans.**  $H_0 : \mu = 52$ ,  $H_1 : \mu = 49 < 52$ , (Left-tailed test).

$Z = -2$ , Not significant.

5. (a) A sample of 450 items is taken from a population whose standard deviation is 20. The mean of the sample is 30. Test whether the sample has come from a population with mean 29. Also calculate the 95% confidence limits for the population mean.

(b) A sample of 400 observations has mean 95 and standard deviation 12. Could it be a random sample from a population with mean 98? What can be the maximum value of the population mean?

6. (a) If the mean age at death of 64 men engaged in an occupation is 52.4 years with standard deviation of 10.2 years, what are the 98% confidence limits for the mean age of all men in that population?

[Calicut Univ. B.Sc. (Subs.), 1989]

(b) The weights of 1500 ball bearings are normally distributed with mean 22.40 and standard deviation 0.048. If 300 random samples of size 36 each are drawn from this population, determine the expected mean and standard deviation of the sampling distribution of means, if sampling is done with replacement.

How many of the random samples in the above problem would have their means between 22.39 and 22.41?

[Madras Univ. B.E., April 1989]

**Hint.**  $E(\bar{X}) = \mu = 22.40$ ;  $S.E.(\bar{X}) = \sigma/\sqrt{n} = 0.048/\sqrt{36} = 0.008$

Required number of samples (out of 300) is :  $300 \times P(22.39 < \bar{X} < 22.41)$

$$= 300 \times P\left(\frac{22.39 - 22.40}{0.008} < Z < \frac{22.41 - 22.40}{0.008}\right); Z \sim N(0, 1)$$

$$= 300 \times P(-1.25 < Z < 1.25) = 600 \times P(0 < Z < 1.25) \approx 237$$

7. (a) A random sample of 500 is drawn from a large number of freshly minted coins. The mean weight of the coins in the sample is 28.57 gm. and the standard deviation is 1.25 gm. What are the limits which have a 49 to 1 chance of including the mean weight of all the coins? How large a sample would have to be drawn to make these limits differ by only 0.1 gm, assuming that the standard deviation of the whole distribution is 1.25 gm.

(b) A research worker wishes to estimate the mean of a population by using sufficiently large sample. The probability is 0.95 that the sample mean will not differ from the true mean of a normal population by more than 25% of the standard deviation. How large a sample should be taken? (Ans.  $n = 62$ .)

8. (a) A normal distribution has mean 0.5 and standard deviation 2.5. Find :

(i) The probability that the mean of a random sample of size 16 from the population is positive.

(ii) The probability that the mean of a sample of size 90 from the population will be negative.

(b) The mean of a certain normal distribution is equal to the standard error of the mean of a random sample of 100 from that distribution. Find the probability, (in terms of an integral), that the mean of a sample of 25 from the distribution will be negative. (Ans. 0.3085.)

(c) The average value  $\bar{x}$  of a random sample of observations from a certain population is normally distributed with mean 20 and standard deviation  $5/\sqrt{n}$ . How large a sample should be drawn in order to have a probability of at least 0.90 that  $\bar{x}$  will lie between 18 and 22. (Karnataka Univ. B.E. 1991)

9. (a) From a population of 169 units it is desired to choose a simple random sample of size  $n$ . If the population standard deviation is 2, determine the smallest ' $n$ ' for which the probability that the sample mean differs from the population mean by more than 0.75 is controlled at 0.05.

(b) An economist would like to estimate the mean income ( $\mu$ ) in a large city. He has decided to use the sample mean as an estimate of  $\mu$  and would like to ensure that the error in estimation is not more than Rs. 100 with probability 0.90. How large a sample should he take if the standard deviation is known to be Rs. 1,000 ? [Delhi Univ. M.A. (Eco.), 1986]

$$\text{Ans. } n = \left[ \frac{z_{\alpha} \cdot \sigma}{E} \right]^2 = \left[ \frac{1.645 \times 1000}{100} \right]^2 = 270.6 \approx 271$$

(c) The management of a manufacturing firm wishes to determine the average time required to complete a certain manual operation. There should be 0.95 confidence that the error in the estimate will not exceed 2 minutes.

What sample size is required if the standard deviation of the time needed to complete the manual operation is estimated by a time and motion study expert as (i) 10 minutes, (ii) 16 minutes ? Explain intuitively (without referring to the formula) why the sample size is large in (ii) than in (i).

(Given  $Z_{0.975} = 1.96$  and  $Z_{0.95} = 1.645$ )

[Delhi Univ. M.C.A., 1987]

$$\text{Ans. (i)} n_1 = \left( \frac{z_{\alpha} \cdot \sigma}{E} \right)^2 = \left( \frac{1.96 \times 10}{2} \right)^2 = 96, \text{ (ii)} n_2 = \left( \frac{1.96 \times 16}{2} \right)^2 = 246.$$

10. (a) Two populations have the same mean, but the standard deviation of one is twice that of the other. Show that in samples of 500 each drawn under simple random conditions, the difference of the means will, in all probability, not exceed  $0.3\sigma$ , where  $\sigma$  is the smaller standard deviation, and assuming the distribution of the difference of the means to be normal, find the probability that it exceeds half that amount. (Ans. 0.1336.)

(b) A simple sample of heights of 6,400 Englishmen has a mean of 67.85 inches and S.D. 2.56 inches, while a simple sample of heights of 1,600 Australians has a mean of 68.55 inches and a S.D. of 2.52 inches. Do the data indicate that Australians are, on the average, taller than Englishmen ?

**Ans.**  $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 < \mu_2, Z = 9.2$ , (significant).

(c) In a random sample of 500, the mean is found to be 20. In another independent sample of 400, the mean is 15. Could the samples have been drawn from the same population with standard deviation 4?

11. (a) The following table presents data on the values of a harvested crop stored in the open and inside a godown:

	Sample size	Mean	$\Sigma (x - \bar{x})^2$
Outside	40	117	8,685
Inside	100	132	27,315

Assuming that the two samples are random and they have been drawn from normal populations with equal variances, examine if the mean value of the harvested crop is affected by weather conditions.

**Ans.**  $Z = 0.342$ ; Not significant.

(b) Samples of students were drawn from two universities and from their weights in kgm., means and standard deviations are calculated. Make a large sample test to test the significance of the difference between the means.

	Mean	S.D.	Size of sample
University A	55	10	400
University B	57	15	100

**Ans.**  $Z = 1.2648$ ; Not significant.

(c) A storekeeper wanted to buy a large quantity of light bulbs from two brands labelled 'one' and 'two'. He bought 100 bulbs from each brand and found by testing that brand 'one' had mean lifetime of 1120 hours and the standard deviation of 75 hours; and brand 'two' had mean lifetime of 1062 hours and standard deviation of 82 hours. Examine whether the difference of means is significant.

12. The mean yield of two sets of plots and their variability are as given below. Examine

(i) whether the difference in the mean yields of the two sets of plots is significant, and

(ii) whether the difference in the variability in yields is significant.

	Set of 40 plots	Set of 60 plots
Mean yield per plot	1258 lb.	1243 lb.
S.D. per plot	34 lb.	28 lb.

**Ans.** (i)  $Z = 2.3$ , (ii)  $Z = 1.3$ .

13. (a) In a survey of incomes of two classes of workers, two random samples gave the following details. Examine whether the differences between the (i) means and (ii) the standard deviations, are significant.

Sample	Size	Mean annual income (in rupees)	Standard deviation (in rupees)
I	100	582	24
II	100	546	28

Examine also whether the first sample could have come from a population with annual mean income of 500 rupees.

(b) The electric light tubes of manufacturer A have a lifetime of 1400 hours, with a standard deviation of 200 hours, while of manufacture B have a mean lifetime of 1200 hours with a standard deviation of 100 hours. If random samples of 125 tubes of each batch are tested, what is the probability that the brand A tubes will have a mean time which is at least (i) 160 hours more than the brand B tubes, and (ii) 250 hours more than the brand B tubes?

**Hint.** Under the assumption of normal population, the sampling distribution of  $(\bar{x}_1 - \bar{x}_2)$  would have mean;  $\mu_1 - \mu_2 = 1400 - 1200 = 200$  hours and standard deviation :

$$\text{S.E. } (\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} = \sqrt{\left\{\frac{(100)^2}{125} + \frac{(200)^2}{125}\right\}} = 20 \text{ hours.}$$

(i) The required probability is given by :

$$\begin{aligned} P\{(\bar{x}_1 - \bar{x}_2) \geq 160\} &= P\left[\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} \geq \frac{160 - 200}{20}\right] \\ &= P(Z \geq -2) = 0.5 + P(-2 < Z < 0) \end{aligned}$$

(ii) The required probability is given by :

$$\begin{aligned} P\{(\bar{x}_1 - \bar{x}_2) \geq 250\} &= P(Z \geq 2.5) = 0.5 - P(0 < Z < 2.5) \\ &= 0.5 - 0.4938 = 0.0062 \end{aligned}$$

14. A random sample of 1,200 men from one State gives the mean pay as Rs. 400 p.m. with a standard deviation of Rs. 60, and a random sample of 1,000 men from another State gives the mean pay as Rs. 500 p.m., with a standard deviation of Rs. 80.

Discuss, (stating clearly the result or theorem used), whether the mean levels of pay of men from the two States differ significantly.

15. (a) A normal population has a mean 0.1 and a standard deviation 2.1. Find the probability that the mean of a sample of size 900 will be negative, it being given that the probability that the absolute value of a standard normal variate exceeds 1.43 is 0.153.

(b) A random sample of 100 articles selected from a batch of 2,000 articles shows that the average diameter of the articles is 0.354 with a standard deviation 0.048. Find 95% confidence interval for the average of this batch of 2,000 articles.

**Hint.** We are given  $n = 100$ ,  $N = 2,000$ ,  $\bar{x} = 0.354$ ,  $s = 0.048$ .

The Standard Error of sample mean  $\bar{x}$  in random sampling from the batch of  $N = 2,000$  is given by : [c.f. (16-23)].

$$\begin{aligned} \text{S.E. } (\bar{x}) &= \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{N-n}{N-1}} \times \frac{s}{\sqrt{n}} \quad (\because \hat{\sigma} = s, \text{ for large } n) \\ &= \sqrt{\frac{2000-100}{2000-1}} \times \frac{0.048}{\sqrt{100}} = 0.00468 \end{aligned}$$

Hence 95% confidence limits for  $\mu$  are given by :

$$\bar{x} \pm 1.96 S.E. (\bar{x}) = 0.354 \pm 1.96 \times 0.00468 = (0.3448, 0.3632)$$

16. (a) Explain the terms :

- (i) Statistic and Parameter
- (ii) Sampling distribution of a statistic, and
- (iii) Standard error of a statistic.

(b) Explain why a random sample of size 30 is to be preferred to a random sample of size 25 to estimate the population mean.

17. (a) Obtain the expressions for the standard error of sampling distributions of : (i) sample mean ( $\bar{X}$ ), and (ii) sample variance ( $S^2$ ), in random sampling from a large population. Assume that  $n$ , the sample size, is large.

(b) Let  $X_1, X_2, \dots, X_n$  be a random sample from a population which has a finite fourth moment  $\mu_r = E(X_i - \mu)^r$ ,  $r = 4$ ;  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2$ ; and

$$\text{let : } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

$$\text{Show that : (i)} S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2,$$

$$(ii) \text{Var}(S^2) = \frac{1}{n} \left[ \mu_4 - \left( \frac{n-3}{n-1} \right) \sigma^4 \right],$$

$$(iii) \text{Cov}(\bar{X}, S^2) = \mu_3/n$$

# **Exact Sampling Distributions (Chi-square Distribution)**

---

**13.1. Chi-Square Variate (Pronounced as Ki - Sky without S).** The square of a standard normal variate is known as a chi-square variate with 1 degree of freedom (d.f.)

Thus if  $X \sim N(\mu, \sigma^2)$ , then  $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$

and  $Z^2 = \left( \frac{X - \mu}{\sigma} \right)^2$ , is a chi-square variate with 1 d.f.

In general, if  $X_i$ , ( $i = 1, 2, \dots, n$ ) are  $n$  independent normal variates with mean  $\mu_i$  and variance  $\sigma_i^2$ , ( $i = 1, 2, \dots, n$ ), then

$$\chi^2 = \sum_{i=1}^n \left( \frac{X_i - \mu_i}{\sigma_i} \right)^2, \text{ is a chi-square variate with } n \text{ d.f.} \quad \dots(13.1)$$

## **13.2. Derivation of the Chi-square Distribution.**

### **First Method—Method of Moment Generating Function.**

If  $X_i$ , ( $i = 1, 2, \dots, n$ ) are independent  $N(\mu_i, \sigma_i^2)$ , we want the distribution of

$$\chi^2 = \sum_{i=1}^n \left( \frac{X_i - \mu_i}{\sigma_i} \right)^2 = \sum_{i=1}^n U_i^2, \text{ where } U_i = \frac{X_i - \mu_i}{\sigma_i}$$

Since  $X_i$ 's are independent,  $U_i$ 's are also independent.

$$M_{\chi^2}(t) = M_{\sum U_i^2}(t) = \prod_{i=1}^n M_{U_i^2}(t) = [M_{U_i^2}(t)]^n,$$

since  $U_i$ 's  $\sim N(0, 1)$  are identically distributed.

Now

$$\begin{aligned} M_{U_i^2}(t) &= E[\exp(tU_i^2)] = \int_{-\infty}^{\infty} \exp(tu_i^2) f(x_i) dx_i \\ &= \int_{-\infty}^{\infty} \exp(tu_i^2) \frac{1}{\sigma_i \sqrt{2\pi}} \exp(-(x_i - \mu_i)^2/2\sigma_i^2) dx_i \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(tu_i^2) \exp(-u_i^2/2) du_i \quad \left[ u_i = \frac{x_i - \mu_i}{\sigma_i} \right] \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left\{ - \left( \frac{1-2t}{2} \right) u_i^2 \right\} du_i \\
 &= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{\left( \frac{1-2t}{2} \right)^{1/2}} = (1-2t)^{-1/2}, \\
 \therefore M_{\chi^2}(t) &= (1-2t)^{-n/2} \quad \dots(13-1a)
 \end{aligned}$$

which is the m.g.f. of a Gamma variate with parameters  $\frac{1}{2}$  and  $\frac{1}{2}n$ .

Hence by uniqueness theorem of m.g.f.'s,

$$\chi^2 = \sum_i^n \left( \frac{X_i - \mu_i}{\sigma_i} \right)^2$$

is a Gamma variate with parameters  $\frac{1}{2}$  and  $\frac{1}{2}n$ .

$$\begin{aligned}
 \therefore dP(\chi^2) &= \frac{(1/2)^{n/2}}{\Gamma(n/2)} \cdot [\exp(-\frac{1}{2}\chi^2)] (\chi^2)^{(n/2)-1} d\chi^2 \\
 &= \frac{1}{2^{n/2} \Gamma(n/2)} [\exp(-\chi^2/2)] (\chi^2)^{(n/2)-1} d\chi^2, \quad 0 \leq \chi^2 < \infty \quad \dots(13-2),
 \end{aligned}$$

which is the required probability density function of chi-square distribution with  $n$  degrees of freedom.

**Remarks 1.** If a random variable  $X$  has a chi-square distribution with  $n$  d.f., we write  $X \sim \chi^2_{(n)}$  and its p.d.f. is given by :

$$f(x) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-x/2} x^{(n/2)-1}; \quad 0 \leq x < \infty \quad \dots(13-2a)$$

2. If  $X \sim \chi^2_{(n)}$ , then  $(X/2) \sim \gamma(n/2)$ .

**Proof.** The p.d.f. of  $Y = \frac{1}{2}X$ , is given by :

$$\begin{aligned}
 g(y) &= f(x) \cdot \left| \frac{dx}{dy} \right| \\
 &= \frac{1}{2^{n/2} \Gamma(n/2)} e^{-y} \cdot (2y)^{(n/2)-1} \cdot 2 \\
 &= \frac{1}{\Gamma(n/2)} e^{-y} y^{(n/2)-1}; \quad 0 \leq y < \infty
 \end{aligned}$$

$$\Rightarrow Y = (X/2) \sim \gamma(n/2)$$

#### Second Method—Method of Induction

If  $X_i$  is a  $N(0, 1)$ , then  $X_i^2/2$  is a  $\gamma(1/2)$  so that  $X_i^2$  is a  $\chi^2$ -variate with d.f. 1.

$$\int_{-\infty}^{\infty} \exp(-a^2 x^2) dx = \frac{\sqrt{\pi}}{a}$$

If  $X_1$  and  $X_2$  are independent standard normal variates then  $X_1^2 + X_2^2$  is a chi-square variate with 2 d.f. which may be proved as follows :

The joint probability differential of  $X_1$  and  $X_2$  is given by :

$$\begin{aligned} dP(x_1, x_2) &= f(x_1, x_2) dx_1 dx_2 = f_1(x_1)f_2(x_2)dx_1 dx_2 \\ &= \frac{1}{2\pi} \exp \left\{ -(x_1^2 + x_2^2)/2 \right\} dx_1 dx_2, -\infty < (x_1, x_2) \leq \infty \end{aligned}$$

Let us now transform to polar co-ordinates by the substitution  $x_1 = r \cos \theta$ ,  $x_2 = r \sin \theta$ . Jacobian of transformation  $J$  is given by

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_2}{\partial r} \\ \frac{\partial x_1}{\partial \theta} & \frac{\partial x_2}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos \theta & \sin \theta \\ -r \sin \theta & r \cos \theta \end{vmatrix} = r$$

Also we have  $r^2 = x_1^2 + x_2^2$  and  $\tan \theta = x_2/x_1$ . As  $x_1$  and  $x_2$  range from  $-\infty$  to  $+\infty$ ,  $r$  varies from 0 to  $\infty$  and  $\theta$  from 0 to  $2\pi$ . The joint probability differential of  $r$  and  $\theta$  now becomes

$$dG(r, \theta) = \frac{1}{2\pi} \exp(-r^2/2) r dr d\theta ; 0 \leq r \leq \infty, 0 \leq \theta \leq 2\pi$$

Integrating over  $\theta$ , the marginal distribution of  $r$  is given by

$$\begin{aligned} dG_1(r) &= \int_0^{2\pi} dG(r, \theta) = r \exp(-r^2/2) dr \left[ \frac{\theta}{2\pi} \right]_0^{2\pi} \\ &= \exp(-r^2/2) r dr \\ \Rightarrow dG_1(r^2) &= \frac{1}{2} \exp(-r^2/2) dr^2 \\ &= \frac{1}{\Gamma(1)} \exp(-r^2/2) (r^2/2)^{1-1} d(r^2/2) \end{aligned}$$

Thus  $\frac{r^2}{2} = \frac{X_1^2 + X_2^2}{2}$  is a  $\gamma(1)$  variate and hence  $r^2 = X_1^2 + X_2^2$  is a  $\chi^2$ -variante with 2 d.f.

For  $n$  variables  $X_i$ , ( $i = 1, 2, \dots, n$ ) we transform  $(X_1, X_2, \dots, X_n)$  to  $(\chi, \theta_1, \theta_2, \dots, \theta_{n-1})$ ; (1 - 1 transformation) by

$$\left. \begin{aligned} x_1 &= \chi \cos \theta_1 \cos \theta_2 \dots \cos \theta_{n-1} \\ x_2 &= \chi \cos \theta_1 \cos \theta_2 \dots \cos \theta_{n-2} \sin \theta_{n-1} \\ x_3 &= \chi \cos \theta_1 \cos \theta_2 \dots \cos \theta_{n-3} \sin \theta_{n-2} \\ &\vdots \\ &\vdots \\ x_j &= \chi \cos \theta_1 \cos \theta_2 \dots \cos \theta_{n-j} \sin \theta_{n-j+1} \\ &\vdots \\ &\vdots \\ x_n &= \chi \sin \theta_1 \end{aligned} \right\} \dots (13.3)$$

where  $\chi > 0$ ,  $-\pi < \theta_1 < \pi$  and  $-\pi/2 < \theta_i < \pi/2$  for  $i = 2, 3, \dots, (n - 1)$ .

$$\text{Then } x_1^2 + x_2^2 + \dots + x_n^2 = \chi^2$$

$$\text{and } |J| = \chi^{n-1} \cos^{n-2} \theta_1 \cos^{n-3} \theta_2 \dots \cos \theta_{n-2}$$

(c.f. Advanced Theory of Statistics Vol 1, by Kendall and Stuart.)

The joint distribution of  $X_1, X_2, \dots, X_n$  viz.,

$$dF(x_1, x_2, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp(-\sum x_i^2/2) \prod_{i=1}^n dx_i$$

transforms to

$$dG(\chi, \theta_1, \theta_2, \dots, \theta_{n-1}) = \exp(-\chi^2/2) \chi^{n-1} \cos^{n-2} \theta_1 \cos^{n-3} \theta_2 \dots \cos \theta_{n-2} d\chi d\theta_1 d\theta_2 \dots d\theta_{n-1}$$

Integrating over  $\theta_1, \theta_2, \dots, \theta_{n-1}$ , we get the distribution of  $\chi^2$  as

$$dP(\chi^2) = k \exp(-\chi^2/2) (\chi^2)^{(n/2)-1} d\chi^2, 0 \leq \chi^2 < \infty$$

The constant  $k$  is determined from the fact that the total probability is unity, i.e.,

$$\int_0^\infty dP(\chi^2) = 1 \Rightarrow k \int_0^\infty \exp(-\chi^2/2) (\chi^2)^{\frac{n}{2}-1} d\chi^2 = 1$$

$$\Rightarrow k = \frac{1}{2^{n/2} \Gamma(n/2)}$$

$$\therefore dP(\chi^2) = \frac{1}{2^{n/2} \Gamma(n/2)} \exp(-\chi^2/2) (\chi^2)^{\frac{n}{2}-1}, 0 \leq \chi^2 < \infty$$

Hence  $\frac{\chi^2}{2} = \frac{1}{2} \sum_{i=1}^n X_i^2$  is a  $\gamma(n/2)$  variate.

$\Rightarrow \chi^2 = \sum_{i=1}^n X_i^2$  is a chi-square variate with  $n$  degrees of freedom

(d.f.) and (13.2) gives p.d.f. of chi-square distribution with  $n$  d.f.

**Remarks 1.** If  $X_i ; i = 1, 2, \dots, n$  are  $n$  independent normal variates with mean  $\mu_i$  and S.D.  $\sigma_i$ , then  $\sum_{i=1}^n \left( \frac{X_i - \mu_i}{\sigma_i} \right)^2$  is a  $\chi^2$ -variante with  $n$  d.f.

**2.** In random sampling from a normal population with mean  $\mu$  and S.D.  $\sigma$ ,  $\bar{x}$  is distributed normally about the mean  $\mu$  with S.D.  $\sigma/\sqrt{n}$ .

$$\therefore \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$\Rightarrow \left[ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right]^2 \text{ is a } \chi^2\text{-variante with 1 d.f.}$$

**3.** Normal distribution is a particular case of  $\chi^2$ -distribution when  $n = 1$ , since for  $n = 1$ ,

$$\begin{aligned} p(\chi^2) &= \frac{1}{\sqrt{2} \Gamma(1/2)} \exp(-\chi^2/2) (\chi^2)^{\frac{1}{2}-1} d\chi^2, \quad 0 \leq \chi^2 < \infty \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\chi^2/2) d\chi, \quad -\infty \leq \chi < \infty \end{aligned}$$

Thus  $\chi$  is a standard normal variate.

**13-3. M.G.F. of  $\chi^2$ -distribution.** Let  $X \sim \chi^2_{(n)}$ , then

$$\begin{aligned} M_X(t) &= E(e^{tx}) = \int_0^\infty e^{tx} f(x) dx \\ &= \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^\infty e^{tx} \cdot e^{-x/2} x^{(n/2)-1} dx \\ &= \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^\infty \exp\left[-\left(\frac{1-2t}{2}\right)x\right] \cdot x^{(n/2)-1} dx \\ &= \frac{1}{2^{n/2} \Gamma(n/2)} \frac{\Gamma(n/2)}{[(1-2t)/2]^{n/2}} \quad [\text{Using Gamma Integral}] \\ &= (1-2t)^{-n/2}, |2t| < 1 \end{aligned} \quad \dots(13-4)$$

which is the required m.g.f. of a  $\chi^2$ -variate with  $n$  d.f.

**Remarks 1.** Using Binomial expansion for negative index, we get from (13-4) if  $|t| < \frac{1}{2}$ .

$$\begin{aligned} M(t) &= 1 + \frac{n}{2} (2t) + \frac{\frac{n}{2} \left(\frac{n}{2} + 1\right)}{2!} (2t)^2 + \dots \\ &\quad + \frac{\frac{n}{2} \left(\frac{n}{2} + 1\right) \left(\frac{n}{2} + 2\right) \dots \left(\frac{n}{2} + r - 1\right)}{r!} (2t)^r + \dots \end{aligned}$$

$$\begin{aligned} \mu'_r &= \text{Coefficient of } \frac{t^r}{r!} \text{ in the expansion of } M(t) \\ &= 2^r \frac{n}{2} \left(\frac{n}{2} + 1\right) \left(\frac{n}{2} + 2\right) \dots \left(\frac{n}{2} + r - 1\right) \\ &= n(n+2)(n+4) \dots (n+2r-2) \end{aligned} \quad \dots(13-4a)$$

**Remark.** If  $n$  is even so that  $n/2$  is a positive integer, then

$$\mu'_r = 2^r \Gamma[(n/2) + r]/\Gamma(n/2) \quad \dots(13-4b)$$

**13-3-1. Cumulant Generating Function of  $\chi^2$ -distribution.** If  $X \sim \chi^2_{(n)}$ , then

$$K_{\chi^2}(t) = \log M_X(t) = -\frac{n}{2} \log(1-2t)$$

$$= \frac{n}{2} \left[ 2t + \frac{(2t)^2}{2} + \frac{(2t)^3}{3} + \frac{(2t)^4}{4} + \dots \right]$$

$\therefore \kappa_1 = \text{Coefficient of } t \text{ in } K(t) = n$

$\kappa_2 = \text{Coefficient of } \frac{t^2}{2!} \text{ in } K(t) = 2n$

$\kappa_3 = \text{Coefficient of } \frac{t^3}{3!} \text{ in } K(t) = 8n$

$\kappa_4 = \text{Coefficient of } \frac{t^4}{4!} \text{ in } K(t) = 48n$

In general,

$$\kappa_r = \text{Coefficient of } \frac{t^r}{r!} \text{ in } K(t) = n 2^{r-1}(r-1)! \quad \dots(13-4c)$$

Hence

$$\begin{aligned} \text{Mean} &= \kappa_1 = n, \text{ Variance} = \mu_2 = \kappa_2 = 2n \\ \mu_3 &= \kappa_3 = 8n, \mu_4 = \kappa_4 + 3\kappa_2^2 = 48n + 12n^2 \\ \beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{8}{n} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{12}{n} + 3 \end{aligned} \quad \left. \right\} \quad \dots(13-4d)$$

**13-3-2. Limiting Form of  $\chi^2$  Distribution for Large Degrees of Freedom.** If  $X \sim \chi^2_{(n)}$ , then  $M_X(t) = (1-2t)^{-n/2}$ ,  $|t| < \frac{1}{2}$

The m.g.f. of standard  $\chi^2$ -variate  $Z$  is given by

$$M_Z(t) = e^{-\mu/\sigma} M_X(t/\sigma) \quad [\mu = n, \sigma^2 = 2n]$$

$$\text{or} \quad M_Z(t) = e^{-\mu/\sigma} (1 - 2t/\sigma)^{-n/2}$$

$$= e^{-n/2\sqrt{2n}} \left( 1 - \frac{2t}{\sqrt{2n}} \right)^{-n/2}$$

$$\begin{aligned} \therefore K_Z(t) &= \log M_Z(t) = -t \sqrt{\frac{n}{2}} - \frac{n}{2} \log \left( 1 - t \sqrt{\frac{2}{n}} \right) \\ &= -t \sqrt{\frac{n}{2}} + \frac{n}{2} \left[ t \cdot \sqrt{\frac{2}{n}} + \frac{t^2}{2} \cdot \frac{2}{n} + \frac{t^3}{3} \left( \frac{2}{n} \right)^{3/2} + \dots \right] \\ &= -t \sqrt{\frac{n}{2}} + t \cdot \sqrt{\frac{n}{2}} + \frac{t^2}{2} + O(n^{-1/2}) \\ &= \frac{t^2}{2} + O(n^{-1/2}), \end{aligned}$$

where  $O(n^{-1/2})$  are terms containing  $n^{1/2}$  and higher powers of  $n$  in the denominator.

$$\therefore \lim_{n \rightarrow \infty} K_Z(t) = \frac{t^2}{2} \Rightarrow M_Z(t) = e^{t^2/2}, \text{ as } n \rightarrow \infty,$$

which is the m.g.f. of a standard normal variate. Hence by uniqueness theorem of m.g.f.  $Z$  is asymptotically normal. In other words, standard  $\chi^2$  variate tends to standard normal variate as  $n \rightarrow \infty$ . Thus,  $\chi^2$ -distribution tends to normal distribution for large d.f.

In practice for  $n \geq 30$ , the  $\chi^2$ -approximation to normal distribution is fairly good. So whenever  $n \geq 30$ , we use the normal probability tables for testing the significance of the value of  $\chi^2$ . That is why in the tables given in the Appendix, the significant values of  $\chi^2$  have been tabulated till  $n = 30$  only.

**Remark.** For the distribution of  $\chi^2$ -variante for large values of  $n$ , see Example 13-7 and also Remark 2 to § 13-7-1.

### 13-3-3. Characteristic Function of $\chi^2$ -distribution.

If  $X \sim \chi^2_{(n)}$ , then

$$\begin{aligned}\phi_X(t) &= E\{\exp(itX)\} = \int_0^\infty \exp(itx) f(x) dx \\ &= \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^\infty \exp\left\{-\left(\frac{1-2it}{2}\right)x\right\} (x)^{\frac{n}{2}-1} dx \\ &= (1-2it)^{-n/2} \quad \dots(13-4e)\end{aligned}$$

### 13-3-4. Mode and skewness of $\chi^2$ -distribution.

Let  $X \sim \chi^2_{(n)}$ , so that

$$f(x) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-x/2} x^{(n/2)-1}, \quad 0 \leq x < \infty \quad \dots(*)$$

Mode of the distribution is the solution of

$$f'(x) = 0 \quad \text{and} \quad f''(x) < 0$$

Logarithmic differentiation w.r.t.  $x$  in (\*) gives :

$$\frac{f'(x)}{f(x)} = 0 - \frac{1}{2} + \left(\frac{n}{2} - 1\right) \cdot \frac{1}{x} = \frac{n-2-x}{2x} \quad \dots(13-5)$$

Since  $f(x) \neq 0$ ,  $f'(x) = 0 \Rightarrow x = n-2$

It can be easily seen that at the point,  $x = (n-2)$ ,  $f''(x) < 0$ .

Hence mode of the chi-square distribution with  $n$  d.f. is  $(n-2)$ .

Also Karl Pearson's coefficient of skewness is given by

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{n-(n-2)}{\sqrt{2n}} = \sqrt{\frac{2}{n}} \quad \dots(13-6)$$

Since Pearson's coefficient of skewness is greater than zero for  $n \geq 1$ , the  $\chi^2$ -distribution is positively skewed. Further since skewness is inversely proportional to the square root of d.f., it rapidly tends to symmetry as the d.f. increases and consequently as  $n \rightarrow \infty$ , the chi-square distribution tends to normal distribution.

**13-3-5. Additive Property of  $\chi^2$ -variates.** *The sum of independent chi-square variates is also a  $\chi^2$ -variante. More precisely, if  $X_i$ , ( $i = 1, 2, \dots, k$ ) are*

If  $n_i$  dependent  $\chi^2$ -variates with  $n_i$  d.f. respectively, then the sum  $\sum_{i=1}^k X_i$  is also a chi-square variate with  $\sum_{i=1}^k n_i$  d.f.

**Proof.** We have

$$M_{X_i}(t) = (1 - 2t)^{-n_i/2}; i = 1, 2, \dots, k.$$

The m.g.f. of the sum  $\sum_{i=1}^k X_i$  is given by

$$\begin{aligned} M_{\sum X_i}(t) &= M_{X_1}(t) M_{X_2}(t) \dots M_{X_k}(t) && [\because X_i \text{'s are independent}] \\ &= (1 - 2t)^{-n_1/2} (1 - 2t)^{-n_2/2} \dots (1 - 2t)^{-n_k/2} \\ &= (1 - 2t)^{-(n_1 + n_2 + \dots + n_k)/2} \end{aligned}$$

which is the m.g.f. of a  $\chi^2$ -variate with  $(n_1 + n_2 + \dots + n_k)$  d.f. Hence by uniqueness theorem of m.g.f.'s,  $\sum_{i=1}^k X_i$  is a  $\chi^2$ -variate with  $\sum_{i=1}^k n_i$  d.f.

**Remarks 1.** Converse is also true, i.e., if  $X_i$ ;  $i = 1, 2, \dots, k$  are  $\chi^2$ -variates with  $n_i$ ;  $i = 1, 2, \dots, k$  d.f. respectively and if  $\sum_{i=1}^k X_i$  is a  $\chi^2$ -variate with  $\sum_{i=1}^k n_i$  d.f., then  $X_i$ 's are independent.

**2.** Another useful version of the converse is as follows :

If  $X$  and  $Y$  are independent non-negative variates such that  $X + Y$  follows chi-square distribution with  $n_1 + n_2$  d.f. and if one of them, say,  $X$  is a  $\chi^2$ -variate with  $n_1$  d.f. then the other, viz.,  $Y$ , is a  $\chi^2$ -variate with  $n_2$  d.f.

**Proof.** Since  $X$  and  $Y$  are independent variates, we have

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) M_Y(t) \\ \Rightarrow (1 - 2t)^{-(n_1 + n_2)/2} &= (1 - 2t)^{-n_1/2} \cdot M_Y(t) \\ &\quad [\because X + Y \sim \chi^2_{(n_1 + n_2)} \text{ and } X \sim \chi^2_{(n_1)}] \\ \Rightarrow M_Y(t) &= (1 - 2t)^{-n_2/2} \end{aligned}$$

which is the m.g.f. of  $\chi^2$ -variate with  $n_2$  d.f. Hence by uniqueness theorem of m.g.f.'s,  $Y \sim \chi^2_{(n_2)}$

**3.** Still another form of the above theorem is "Cochran theorem", which is as follows :

Let  $X_1, X_2, \dots, X_n$  be independently distributed as standard normal variates, i.e.,  $N(0, 1)$ . Let

$$\sum_{i=1}^n X_i^2 = Q_1 + Q_2 + \dots + Q_k$$

where each  $Q_i$  is a sum of squares of linear combinations of  $X_1, X_2, \dots, X_n$  with  $n_i$  degrees of freedom. Then if  $n_1 + n_2 + \dots + n_k = n$ , the quantities  $Q_1, Q_2, \dots, Q_k$  are independent  $\chi^2$ -variates with  $n_1, n_2, \dots, n_k$  d.f. respectively.

**13.4. Chi-square Probability Curve.** We get from (13.5)

$$f'(x) = \left[ \frac{n - 2 - x}{2x} \right] f(x). \quad \dots(13.7)$$

Since  $x > 0$  and  $f(x)$  being p.d.f. is always non-negative, we get from (13.7) :

$$f'(x) < 0 \text{ if } (n - 2) \leq 0,$$

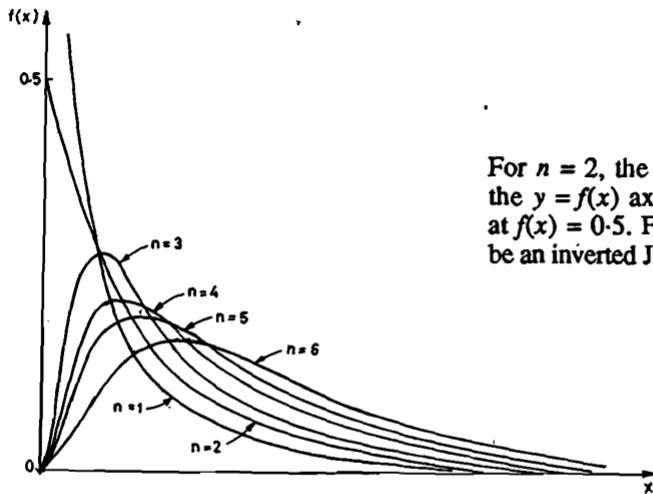
for all values of  $x$ . Thus the  $\chi^2$ -probability curve for 1 and 2 degrees of freedom is monotonically decreasing.

When  $n > 2$ ,

$$f'(x) = \begin{cases} > 0, & \text{if } x < (n - 2) \\ = 0, & \text{if } x = n - 2 \\ < 0, & \text{if } x > (n - 2) \end{cases}$$

This implies that for  $n > 2$ ,  $f(x)$  is monotonically increasing for  $0 < x < (n - 2)$  and monotonically decreasing for  $(n - 2) < x < \infty$ , while at  $x = n - 2$ , it attains the maximum value.

For  $n \geq 1$ , as  $x$  increases,  $f(x)$  decreases rapidly and finally tends to zero as  $x \rightarrow \infty$ . Thus for  $n > 1$ , the  $\chi^2$ -probability curve is positively skewed [c.f. (13.6)] towards higher values of  $x$ . Moreover,  $x$ -axis is an asymptote to the curve. The shape of the curve for  $n = 1, 2, 3, \dots, 6$  is given below.



For  $n = 2$ , the curve will meet the  $y = f(x)$  axis at  $x = 0$ , i.e., at  $f(x) = 0.5$ . For  $n = 1$ , it will be an inverted J-shaped curve.

PROBABILITY CURVE OF CHI-SQUARE DISTRIBUTION

**Theorem 13.1.** If  $\chi_1^2$  and  $\chi_2^2$  are two independent  $\chi^2$ -variates with  $n_1$  and  $n_2$  d.f. respectively, then

$\frac{\chi_1^2}{\chi_2^2}$  is a  $\beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$  variate.

(Gauhati Univ. M.Sc., 1992)

**Proof.** Since  $\chi_1^2$  and  $\chi_2^2$  are independent  $\chi^2$ -variates with  $n_1$  and  $n_2$  d.f. respectively, their joint probability differential is given by the compound probability theorem as

$$\begin{aligned} dP(\chi_1^2, \chi_2^2) &= dP_1(\chi_1^2) dP_2(\chi_2^2) \\ &= \left[ \frac{1}{2^{n_1/2} \Gamma(n_1/2)} \exp(-\chi_1^2/2) (\chi_1^2)^{(n_1/2)-1} d\chi_1^2 \right] \\ &\quad \times \left[ \frac{1}{2^{n_2/2} \Gamma(n_2/2)} \exp(-\chi_2^2/2) (\chi_2^2)^{(n_2/2)-1} d\chi_2^2 \right] \\ &= \frac{1}{2^{(n_1+n_2)/2} \Gamma(n_1/2) \Gamma(n_2/2)} \exp\left(-(\chi_1^2 + \chi_2^2)/2\right) \\ &\quad \times (\chi_1^2)^{\frac{n_1}{2}-1} (\chi_2^2)^{\frac{n_2}{2}-1} d\chi_1^2 d\chi_2^2, \quad 0 \leq (\chi_1^2, \chi_2^2) < \infty \end{aligned}$$

Let us make the transformation :

$$\begin{array}{lll} u = \chi_1^2/\chi_2^2 & \text{and} & v = \chi_2^2 \\ \text{so that } \chi_1^2 = uv & \text{and} & \chi_2^2 = v \end{array}$$

Jacobian of transformation  $J$  is given by

$$J = \frac{\partial(\chi_1^2, \chi_2^2)}{\partial(u, v)} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v$$

Thus the joint distribution of random variables  $U$  and  $V$  becomes

$$\begin{aligned} dG(u, v) &= \frac{1}{2^{(n_1+n_2)/2} \Gamma(n_1/2) \Gamma(n_2/2)} \exp\left\{-\left(1+\frac{1}{v}\right)v/2\right\} \\ &\quad \times (uv)^{\frac{n_1}{2}-1} v^{\frac{n_2}{2}-1} |J| du dv, \\ &= \frac{1}{2^{(n_1+n_2)/2} \Gamma(n_1/2) \Gamma(n_2/2)} \exp\left\{-\left(1+\frac{1}{v}\right)v/2\right\} \\ &\quad \times u^{\frac{n_1}{2}-1} v^{\frac{n_1+n_2}{2}-1} du dv, \quad 0 \leq (u, v) < \infty \end{aligned}$$

Integrating w.r.t.  $v$  over the range 0 to  $\infty$ , we get the marginal distribution

$$\text{of } U \text{ as : } dG_1(u) = \int_0^\infty dG(u, v)$$

$$\begin{aligned} &= \frac{1}{2^{(n_1+n_2)/2} \Gamma(n_1/2) \Gamma(n_2/2)} u^{(n_1/2)-1} du \\ &\quad \times \int_0^\infty \exp\left\{-\left(\frac{1+u}{2}\right)v\right\} v^{\{(n_1+n_2)/2\}-1} dv \end{aligned}$$

$$\begin{aligned}
 &= \frac{u^{(n_1/2)-1}}{2^{(n_1+n_2)/2} \Gamma(n_1/2) \Gamma(n_2/2)} \cdot \frac{\Gamma((n_1+n_2)/2)}{[(1+u)/2]^{(n_1+n_2)/2}} du \\
 &= \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{u^{(n_1/2)-1}}{[1+u]^{(n_1+n_2)/2}} du, \quad 0 \leq u < \infty
 \end{aligned}$$

Hence  $U = \frac{\chi_1^2}{\chi_2^2}$  is a  $\beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$  variate.

**Theorem 13.2.** If  $\chi_1^2$  and  $\chi_2^2$  are independent  $\chi^2$ -variates with  $n_1$  and  $n_2$  d.f. respectively, then

$$U = \frac{\chi_1^2}{\chi_1^2 + \chi_2^2} \text{ and } V = \chi_1^2 + \chi_2^2$$

are independently distributed,  $U$  as a  $\beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$  variate and  $V$  as a  $\chi^2$  variate with  $(n_1+n_2)$  d.f.

**Proof.** As the Theorem 13.1, we have

$$\begin{aligned}
 dP(\chi_1^2, \chi_2^2) &= \frac{1}{2^{(n_1+n_2)/2} \Gamma(n_1/2) \Gamma(n_2/2)} \exp\{-(\chi_1^2 + \chi_2^2)/2\} \\
 &\quad \times (\chi_1^2)^{(n_1/2)-1} (\chi_2^2)^{(n_2/2)-1} d\chi_1^2 d\chi_2^2, \quad 0 \leq (\chi_1^2, \chi_2^2) < \infty
 \end{aligned}$$

Let us transform to  $u$  and  $v$  defined as follows :

$$u = \frac{\chi_1^2}{\chi_1^2 + \chi_2^2} \text{ and } v = \chi_1^2 + \chi_2^2$$

so that  $\chi_1^2 = uv$  and  $\chi_2^2 = v - \chi_1^2 = (1-u)v$

As  $\chi_1^2$  and  $\chi_2^2$  both range from 0 to  $\infty$ ;  $u$  ranges from 0 to 1 and  $v$  from 0 to  $\infty$ .

Jacobian of transformation  $J$  is

$$J = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

$$\begin{aligned}
 \therefore dG(u, v) &= \frac{1}{2^{(n_1+n_2)/2} \Gamma(n_1/2) \Gamma(n_2/2)} \exp(-v/2) (uv)^{(n_1/2)-1} \\
 &\quad \times [(1-u)v]^{(n_2/2)-1} |J| du dv \\
 &= \frac{1}{2^{(n_1+n_2)/2} \Gamma(n_1/2) \Gamma(n_2/2)} u^{(n_1/2)-1} (1-u)^{(n_2/2)-1} \\
 &\quad \times \exp(-v/2) \cdot v^{((n_1+n_2)/2)-1} du dv \\
 &= \left[ \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} u^{(n_1/2)-1} (1-u)^{(n_2/2)-1} du \right] \\
 &\quad \times \left[ \frac{1}{2^{(n_1+n_2)/2} \Gamma((n_1+n_2)/2)} \exp(-v/2) v^{((n_1+n_2)/2)-1} dv \right]
 \end{aligned}$$

Since the joint probability differential of  $U$  and  $V$  is the product of their respective probability differentials,  $U$  and  $V$  are independently distributed, with

$$dG_1(u) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} u^{(n_1/2)-1} (1-u)^{(n_2/2)-1} du, \quad 0 \leq u \leq 1$$

and

$$dG_2(v) = \frac{1}{2^{(n_1+n_2)/2} \Gamma((n_1+n_2)/2)} \exp(-v/2) v^{\{(n_1+n_2)/2\}-1} dv, \quad 0 \leq v < \infty$$

i.e.,  $U$  as a  $\beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$  variate and  $V$  as a  $\chi^2$ -variate with  $(n_1 + n_2)$  d.f

**Remark.** The results in Theorems 13-1 and 13-2 can be summarised as follows :

If  $X \sim \chi^2_{(n_1)}$  and  $Y \sim \chi^2_{(n_2)}$  are independent chi-square variates then :

(i)  $X + Y \sim \chi^2_{(n_1 + n_2)}$  i.e., the sum of two independent chi-square variates is also a chi-square variate.

(ii)  $\frac{X}{Y} \sim \beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$  i.e., the ratio of two independent chi-square variates is a  $\beta_2$ -variate.

(iii)  $\frac{X}{X+Y} \sim \beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$

**Theorem 13-3.** In a random and large sample,

$$\chi^2 = \sum_{i=1}^k \left[ \frac{(n_i - np_i)^2}{np_i} \right], \quad \dots(13-8)$$

follows chi-square distribution approximately with  $(k - 1)$  degrees of freedom, where  $n_i$  is the observed frequency and  $np_i$  is the corresponding expected frequency of the  $i$ th class, ( $i = 1, 2, \dots, k$ ),  $\sum_{i=1}^k n_i = n$ .

**Proof.** Let us consider a random sample of size  $n$ , whose members are distributed at random in  $k$  classes or cells. Let  $p_i$  be the probability that sample observation will fall in the  $i$ th cell, ( $i = 1, 2, \dots, k$ ). Then the probability  $P$  of there being  $n_i$  members in the  $i$ th cell, ( $i = 1, 2, \dots, k$ ) respectively is given by the multinomial probability law, by the expression

$$P = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k},$$

where  $\sum_{i=1}^k n_i = n$  and  $\sum_{i=1}^k p_i = 1$ .

If  $n$  is sufficiently large so that  $n_i$ , ( $i = 1, 2, \dots, k$ ) are not small then using Stirling's approximation to factorials for large  $n$ , viz.,

$$\lim_{n \rightarrow \infty} (n!) \approx \sqrt{2\pi} e^{-n} n^{n + \frac{1}{2}}, \text{ we get}$$

$$\begin{aligned} P &= \frac{\sqrt{2\pi} e^{-n} n^{n + \frac{1}{2}}}{(\sqrt{2\pi})^k e^{-(n_1 + n_2 + \dots + n_k)}} \times \frac{p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}}{n_1^{n_1 + \frac{1}{2}} n_2^{n_2 + \frac{1}{2}} \dots n_k^{n_k + \frac{1}{2}}} \\ &\approx \frac{e^{-n} n^{n + \frac{1}{2}} \left(\frac{np_1}{n_1}\right)^{n_1 + \frac{1}{2}} \left(\frac{np_2}{n_2}\right)^{n_2 + \frac{1}{2}} \dots \left(\frac{np_k}{n_k}\right)^{n_k + \frac{1}{2}}}{(\sqrt{2\pi})^{k-1} e^{-n} n^{n_1 + n_2 + \dots + n_k + (k/2)} (p_1 p_2 \dots p_k)^{1/2}} \\ &\approx C \prod_{i=1}^k \left(\frac{np_i}{n_i}\right)^{n_i + \frac{1}{2}} \end{aligned}$$

$$\text{where } C = \frac{1}{(2\pi)^{(k-1)/2} n^{(k-1)/2} (p_1 p_2 \dots p_k)^{1/2}},$$

is a constant independent of  $n_i$ 's.

$$\therefore \log P \approx \log C + \sum_{i=1}^k (n_i + \frac{1}{2}) \log \left(\frac{np_i}{n_i}\right)$$

$$\Rightarrow \log(P/C) \approx \sum_{i=1}^k (n_i + \frac{1}{2}) \log \left(\frac{\lambda_i}{n_i}\right), \quad \dots(*)$$

where  $\lambda_i = np_i$  is the expected frequency for the  $i$ th cell, i.e.,

$$E(n_i) = np_i = \lambda_i, \quad (i = 1, 2, \dots, k).$$

Let us define

$$\xi_i = \frac{n_i - \lambda_i}{\sqrt{\lambda_i}},$$

$$\text{so that } n_i - \lambda_i = \xi_i \sqrt{\lambda_i} \Rightarrow n_i = \lambda_i + \xi_i \sqrt{\lambda_i} \quad \dots(**)$$

Substituting in (\*), we get

$$\begin{aligned} \log(P/C) &\approx \sum_{i=1}^k (\lambda_i + \xi_i \sqrt{\lambda_i} + \frac{1}{2}) \log \left[ \frac{\lambda_i}{\lambda_i + \xi_i \sqrt{\lambda_i}} \right] \\ &= \sum_{i=1}^k (\lambda_i + \xi_i \sqrt{\lambda_i} + \frac{1}{2}) \log [1/(1 + \xi_i / \sqrt{\lambda_i})] \\ &= - \sum_{i=1}^k (\lambda_i + \xi_i \sqrt{\lambda_i} + \frac{1}{2}) \log (1 + (\xi_i / \sqrt{\lambda_i})) \end{aligned}$$

If we assume that  $\xi_i$  is small compared with  $\lambda_i$ , the expansion of  $\log 1 + (\xi_i / \sqrt{\lambda_i})$  in ascending powers of  $\xi_i / \sqrt{\lambda_i}$  is valid.

$$\begin{aligned}\therefore \log P/C &\approx - \sum_{i=1}^k (\lambda_i + \xi_i \sqrt{\lambda_i} + \frac{1}{2} \lambda_i) \left[ \frac{\xi_i}{\sqrt{\lambda_i}} - \frac{1}{2} \frac{\xi_i^2}{\lambda_i} + O(1/\lambda_i^{3/2}) \right] \\ &\approx - \sum_{i=1}^k \left[ \xi_i \sqrt{\lambda_i} - \frac{1}{2} \xi_i^2 + \xi_i^2 + O(\lambda_i^{-1/2}) \right],\end{aligned}$$

neglecting higher powers of  $\xi_i/\sqrt{\lambda_i}$  if  $\xi_i$  is small compared with  $\lambda_i$ .

Since  $n$  is large, so is  $\lambda_i = np_i$ . Hence  $O(\lambda_i^{-1/2}) \rightarrow 0$  for large  $n$ .

$$\begin{aligned}\text{Also } \sum_{i=1}^k \xi_i \sqrt{\lambda_i} &= \sum_{i=1}^k (n_i - \lambda_i) = \sum_{i=1}^k n_i - \sum_{i=1}^k \lambda_i \\ &= \sum_{i=1}^k n_i - n \sum_{i=1}^k p_i = n - n = 0 \quad (\because \sum n_i = n, \sum p_i = 1)\end{aligned}$$

$$\begin{aligned}\therefore \log(P/C) &\approx - \left[ \sum_{i=1}^k \xi_i \sqrt{\lambda_i} + \frac{1}{2} \sum_{i=1}^k \xi_i^2 + O(\lambda_i^{-1/2}) \right] \approx - \frac{1}{2} \sum_{i=1}^k \xi_i^2 \\ \Rightarrow \quad P &\approx C \exp \left( - \frac{1}{2} \sum_{i=1}^k \xi_i^2 \right).\end{aligned}$$

which shows that  $\xi_i$ , ( $i = 1, 2, \dots, k$ ) are distributed as independent standard normal variates.

$$\text{Hence } \sum_{i=1}^k \xi_i^2 = \sum_{i=1}^k \left[ \frac{(n_i - \lambda_i)^2}{\lambda_i} \right],$$

being the sum of the squares of  $k$  independent standard normal variates is a  $\chi^2$ -variate with  $(k-1)$  d.f., one d.f. being lost because of the linear constraint

$$\sum_{i=1}^k \xi_i \sqrt{\lambda_i} = \sum (n_i - \lambda_i) = 0 \Rightarrow \sum_{i=1}^k n_i = \sum_{i=1}^k \lambda_i \quad \dots (***)$$

**Remarks 1.** If  $O_i$  and  $E_i$  ( $i = 1, 2, \dots, k$ ), be a set of observed and expected frequencies, then

$$\chi^2 = \sum_{i=1}^k \left[ \frac{(O_i - E_i)^2}{E_i} \right], \quad \left( \sum_{i=1}^k O_i = \sum_{i=1}^k E_i \right) \quad \dots (13-8a)$$

follows chi-square distribution with  $(k-1)$  d.f.

Another convenient form of this formula is as follows :

$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \left( \frac{O_i^2 + E_i^2 - 2O_iE_i}{E_i} \right) = \sum_{i=1}^k \left( \frac{O_i^2}{E_i} + E_i - 2O_i \right) \\ &= \sum_{i=1}^k \left( \frac{O_i^2}{E_i} \right) + \sum_{i=1}^k E_i - 2 \sum_{i=1}^k O_i\end{aligned}$$

$$= \sum_{i=1}^k \left( \frac{O_i^2}{E_i} \right) - N, \quad \dots(13.8b)$$

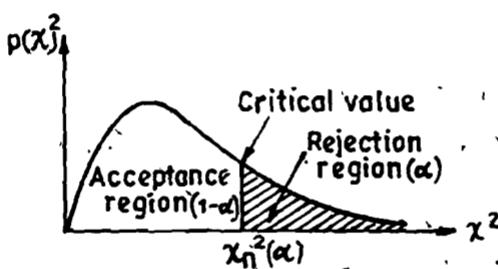
where  $\sum_{i=1}^k O_i = \sum_{i=1}^k E_i = N$  (say), is the total frequency.

2. Conditions for the Validity of  $\chi^2$ -test.  $\chi^2$ -test is an approximate test for large values of  $n$ . For the validity of chi-square test of 'goodness of fit' between theory and experiment, the following conditions must be satisfied :

- (i) The sample observations should be independent.
  - (ii) Constraints on the cell frequencies, if any, should be linear, e.g.,  $\sum n_i = \sum \lambda_i$  or  $\sum O_i = \sum E_i$ .
  - (iii)  $N$ , the total frequency should be reasonably large, say, greater than 50.
  - (iv) No theoretical cell frequency should be less than 5. (The chi square distribution is essentially a continuous distribution but it cannot maintain its character of continuity if cell frequency is less than 5). If any theoretical cell frequency is less than 5, then for the application of  $\chi^2$ -test, it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f. lost in pooling.
3. It may be noted that the  $\chi^2$ -test depends only on the set of observed and expected frequencies and on degrees of freedom (d.f.). It does not make any assumptions regarding the parent population from which the observations are taken. Since  $\chi^2$  defined in (13.8) does not involve any population parameters, it is termed as a statistic and the test is known as *Non-Parametric Test* or *Distribution-Free Test*.

4. Critical Values. Let  $\chi_n^2(\alpha)$  denote the value of chi-square for  $n$  d.f. such that the area to the right of this point is  $\alpha$ , i.e.,

$$P[\chi^2 > \chi_n^2(\alpha)] = \alpha \quad \dots(13.8c)$$



The value  $\chi_n^2(\alpha)$  defined in (13.8c) is known as the *upper (right-tailed) α-point or Critical Value or Significant Value of chi-square* for  $n$  d.f. and has been tabulated for different values of  $n$  and  $\alpha$  in Table VI in the Appendix at the end of the book. From these tables we observe that the critical values of  $\chi^2$  increase as  $n$  (d.f.) increases and level of significance ( $\alpha$ ) decreases.

The critical values for left-tailed test or two tailed tests can be obtained from the above table, as discussed in Remark 1 to § 16.7.4.

**13.6. Linear Transformation.** Let us suppose that the given set of variables  $\mathbf{X}' = (x_1, x_2, \dots, x_n)$  is transformed to a new set of variables  $\mathbf{Y}' = (y_1, y_2, \dots, y_n)$  by means of the linear transformation :

$$\left. \begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \\ &\vdots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n \end{aligned} \right\} \quad \dots(13.9)$$

i.e.,  $y_i = a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n ; i = 1, 2, \dots, n$

In matrix notation, this system of linear equations can be expressed symbolically as

$$\mathbf{Y} = \mathbf{AX} \quad \dots(13.10)$$

where  $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$

From matrix theory, we know that the system (13.10) has a unique solution iff  $|\mathbf{A}| \neq 0$ . In other words, we can express  $\mathbf{X}$  uniquely in terms of  $\mathbf{Y}$  if  $\mathbf{A}$  is non-singular and the solution is given by

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y} \quad \dots(13.10a)$$

where  $\mathbf{A}^{-1}$  is the inverse of the square matrix  $\mathbf{A}$ .

The linear transformation defined in (13.9) or (13.10) is said to be *orthogonal* if

$$\mathbf{X}'\mathbf{X} = \mathbf{Y}'\mathbf{Y} \quad \dots(13.11)$$

$$\Rightarrow \mathbf{X}'\mathbf{X} = (\mathbf{AX})' \mathbf{AX} = \mathbf{X}'(\mathbf{A}'\mathbf{A})\mathbf{X}$$

$$\Rightarrow \mathbf{A}'\mathbf{A} = \mathbf{I}_n \quad \dots(13.11a)$$

$\Rightarrow \mathbf{A}$  is an orthogonal matrix.

More elaborately

$$\mathbf{X}'\mathbf{X} = \mathbf{Y}'\mathbf{Y}$$

$$\Rightarrow \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2 = \sum_{i=1}^n (a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n)^2, \quad \dots(*)$$

for every set of variables,  $(x_1, x_2, \dots, x_n)$ .

If we write  $\delta_{ij} = \sum_{k=1}^n a_{ik}a_{kj}$ ,  $(i, j = 1, 2, \dots, n)$ ,

then (\*) implies that  $\delta_{ij}$  is a Kronecker delta so that

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad \dots(13.11b)$$

whence it follows that  $\mathbf{A}$  is an orthogonal matrix.

**Linear Orthogonal Transformation.** Def.: A linear transformation  $\mathbf{Y} = \mathbf{AX}$ , is said to be orthogonal if  $\mathbf{A}$  is an orthogonal matrix.

**Remarks 1.** It is very easy to verify the equivalence of the following two definitions of an orthogonal matrix.

**Def. 1.** A square matrix  $\mathbf{A}$  ( $n \times n$ ) is said to be orthogonal if  $\mathbf{A}'\mathbf{A} = \mathbf{AA}' = \mathbf{I}_n$ .

**Def. 2.** A square matrix  $\mathbf{A}$  is said to be orthogonal if the transformation  $\mathbf{Y} = \mathbf{AX}$  transforms  $\mathbf{X}'\mathbf{X}$  to  $\mathbf{Y}'\mathbf{Y}$ .

2. If  $\mathbf{Y} = \mathbf{AX}$  is an orthogonal transformation, then  $\mathbf{Y}'\mathbf{Y} = \mathbf{X}'\mathbf{X}$  and  $\mathbf{A}'\mathbf{A} = \mathbf{AA}' = \mathbf{I}_n$ .

**Theorem 13.4. (Fisher's Lemma).** If  $X_i$ , ( $i = 1, 2, \dots, n$ ) are independent  $N(0, \sigma^2)$  and they are transformed to a new set of variables  $Y_i$ , ( $i = 1, 2, \dots, n$ ), by means of a linear orthogonal transformation, then  $Y_i$ , ( $i = 1, 2, \dots, n$ ) are also independent  $N(0, \sigma^2)$ .

**Proof.** Let the linear orthogonal transformation be

$$\mathbf{Y} = \mathbf{AX} \text{ so that } \mathbf{Y}'\mathbf{Y} = \mathbf{X}'\mathbf{X} \text{ and } \mathbf{A}'\mathbf{A} = \mathbf{I}_n$$

Since  $X_i$ , ( $i = 1, 2, \dots, n$ ) are independent  $N(0, \sigma^2)$ , their joint density function is given by

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left(-\sum_{i=1}^n x_i^2/2\sigma^2\right), -\infty < (x_1, x_2, \dots, x_n) < \infty \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp(-X'X/2\sigma^2) \end{aligned}$$

The joint density of  $(Y_1, Y_2, \dots, Y_n)$  becomes

$$g(y_1, y_2, \dots, y_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp(-Y'Y/2\sigma^2) |J|$$

where  $\frac{1}{J} = \frac{\partial(y_1, y_2, \dots, y_n)}{\partial(x_1, x_2, \dots, x_n)} = |\mathbf{A}|$

Now  $\mathbf{A}'\mathbf{A} = \mathbf{I}_n$

$\Rightarrow |\mathbf{A}'\mathbf{A}| = |\mathbf{I}_n| = 1$

$\Rightarrow |\mathbf{A}'| |\mathbf{A}| = 1$

$\Rightarrow |\mathbf{A}|^2 = 1 \quad (\because |\mathbf{A}'| = |\mathbf{A}|)$

$\Rightarrow |\mathbf{A}| = \pm 1$

$\therefore |J| = |\pm 1| = 1$

$$\therefore g(y_1, y_2, \dots, y_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp(-Y'Y/2\sigma^2)$$

$$= \prod_{i=1}^n \left[ \frac{1}{\sigma\sqrt{2\pi}} \exp(-y_i^2/2\sigma^2) \right]$$

Hence  $Y_i$ , ( $i = 1, 2, \dots, n$ ) are independent  $N(0, \sigma^2)$ .

**Theorem 13-5.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal population with mean  $\mu$  and variance  $\sigma^2$ . Then

$$(i) \quad \bar{X} \sim N(\mu, \sigma^2/n),$$

$$(ii) \quad \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \text{ is a } \chi^2\text{-variate with } (n-1) \text{ d.f., and}$$

$$(iii) \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ and } \frac{n s^2}{\sigma^2} = \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \text{ are independently distributed.}$$

[Delhi Univ. B.Sc. (Maths Hons.) 1987;  
Sardar Patel Univ. B.Sc. 1992]

**Proof.** The joint probability differential of  $X_1, X_2, \dots, X_n$  is given by

$$dP(x_1, x_2, \dots, x_n) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] dx_1 dx_2 \dots dx_n ;$$

$$-\infty < (x_1, x_2, \dots, x_n) < \infty$$

Let us transform to the variables  $Y_i$ , ( $i = 1, 2, \dots, n$ ) by means of a linear orthogonal transformation ( $\mathbf{Y} = \mathbf{AX}$ ) (c.f. § 13-6, page 13-16).

Let us choose in particular

$$a_{11} = a_{12} = \dots = a_{1n} = 1/\sqrt{n}$$

$$\Rightarrow y_1 = \frac{1}{\sqrt{n}} (x_1 + x_2 + \dots + x_n) = \sqrt{n} \bar{x} \quad \dots (*)$$

(It can be easily seen that the above choice of  $a_{11}, a_{12}, \dots, a_{1n}$  satisfies the condition of orthogonality; viz.,  $\sum_{i=1}^n a_{ij}^2 = 1$ ).

Since the transformation is orthogonal, we have

$$\begin{aligned} \sum_{i=1}^n y_i^2 &= \sum_{i=1}^n x_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n \bar{x}^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + y_1^2 \quad [\text{From } (*)] \end{aligned}$$

$$\Rightarrow \sum_{i=2}^n y_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \dots (**)$$

$$\begin{aligned} \text{Also} \quad \sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \\ &= \sum_{i=2}^n y_i^2 + n(\bar{x} - \mu)^2 \quad [\text{From } (**)] \end{aligned}$$

As in Theorem 13-4, the Jacobian of transformation  $J = \pm 1$ .

Thus the joint density function of  $X_1, X_2, \dots, X_n$  transforms to

$$\begin{aligned}
 dG(y_1, y_2, \dots, y_n) &= \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \left\{ \sum_{i=2}^n y_i^2 + n(\bar{x} - \mu)^2 \right\} \right] \\
 &\quad \times |J| dy_1 dy_2 \dots dy_n \\
 &= \left[ \frac{1}{\sqrt{2\pi}(\sigma/\sqrt{n})} \exp \left\{ -\frac{n}{2\sigma^2} (\bar{x} - \mu)^2 \right\} d\bar{x} \right] \\
 &\quad \times \left[ \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^{n-1} \exp \left\{ -\sum_{i=2}^n \frac{y_i^2}{2\sigma^2} \right\} dy_2 dy_3 \dots dy_n \right] \\
 &\quad (\because dy_1 = \sqrt{n} d\bar{x})
 \end{aligned}$$

Thus  $\bar{X}$  and  $\sum_{i=2}^n Y_i^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = ns^2$ , (where  $s^2$  is the sample variance), are independently distributed, which establishes part (iii) of the Theorem. Moreover  $\bar{X} \sim N(\mu, \sigma^2/n)$  and  $Y_i$ , ( $i = 1, 2, 3, \dots, n$ ) are independent  $N(0, \sigma^2)$ . Hence

$$\sum_{i=2}^n \frac{Y_i^2}{\sigma^2} = \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2,$$

being the sum of squares of  $(n-1)$  independent standard normal variates is distributed as  $\chi^2$ -variate with  $(n-1)$  d.f.

**Aliter.** The alternative proof of the above Theorem is based on the use of m.g.f.'s and is given below.

We shall first prove that :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad X_i - \bar{X}, \quad i = 1, 2, \dots, n \quad \dots(i)$$

are independently distributed.

The joint m.g.f. of  $\bar{X}$  and  $(X_i - \bar{X})$  is given by :

$$\begin{aligned}
 M(t_1, t_2) &= E \left[ e^{t_1 \bar{X} + t_2 (X_i - \bar{X})} \right] = E \left[ e^{(t_1 - t_2) \bar{X} + t_2 X_i} \right] \\
 &= E \left[ \exp \left\{ \frac{t_1 - t_2}{n} \cdot \sum_{i=1}^n X_i + t_2 X_i \right\} \right] \\
 &= E \left[ \exp \left\{ \left( \frac{t_1 - t_2}{n} + t_2 \right) X_i + \frac{t_1 - t_2}{n} \sum_{j=1, j \neq i}^n X_j \right\} \right]
 \end{aligned}$$

$$= E \left[ \exp \left\{ \left( \frac{t_1 - t_2}{n} + t_2 \right) X_i \right\} \right] \cdot E \left[ \exp \left( \frac{t_1 - t_2}{n} \sum_{\substack{j=1 \\ (j \neq i)}}^n X_j \right) \right] \quad \dots(ii)$$

( $\because X_1, X_2, \dots, X_n$  are independent)

Now  $U = \sum_{\substack{j=1 \\ (j \neq i)}}^n X_j$ , being the sum of  $(n-1)$  i.i.d.  $N(\mu, \sigma^2)$  variates

is a  $N\{(n-1)\mu, (n-1)\sigma^2\}$  variate.

$$\therefore M_U(t) = \exp [t \cdot (n-1) \mu + t^2 \cdot (n-1) \sigma^2/2] \quad \dots(iii)$$

Hence

$$\begin{aligned} E \left[ \exp \left\{ \frac{t_1 - t_2}{n} \sum_{\substack{j=1 \\ (j \neq i)}}^n X_j \right\} \right] &= E \left[ \exp \left( \frac{t_1 - t_2}{n} \cdot U \right) \right] \\ &= M_U[(t_1 - t_2)/n] \\ &= \exp \left[ \left( \frac{t_1 - t_2}{n} \right) (n-1) \mu + \left( \frac{t_1 - t_2}{n} \right)^2 (n-1) \frac{\sigma^2}{2} \right] \\ &\quad [\text{On using (iii)}] \end{aligned} \quad \dots(iv)$$

$$\begin{aligned} \text{and } E \left[ \exp \left\{ \left( \frac{t_1 - t_2}{n} + t_2 \right) X_i \right\} \right] &= M_{X_i} \left( \frac{t_1 - t_2}{n} + t_2 \right) \\ &= \exp \left[ \left( \frac{t_1 - t_2}{n} + t_2 \right) \mu + \left( \frac{t_1 - t_2}{n} + t_2 \right)^2 \frac{\sigma^2}{2} \right] \\ &\quad [\because X_i \sim N(\mu, \sigma^2)] \end{aligned} \quad \dots(v)$$

Substituting from (iv) and (v) in (ii), we get

$$\begin{aligned} M(t_1, t_2) &= \exp \left[ \left\{ \left( \frac{t_1 - t_2}{n} \right) (n-1) + \left( \frac{t_1 - t_2}{n} + t_2 \right) \right\} \mu \right] \\ &\quad \times \exp \left[ \left\{ \left( \frac{t_1 - t_2}{n} \right)^2 \cdot (n-1) + \left( \frac{t_1 - t_2}{n} + t_2 \right)^2 \right\} \frac{\sigma^2}{2} \right] \\ &= \exp \left[ t_1 \mu + \frac{1}{2} t_1^2 \cdot \left( \frac{\sigma^2}{n} \right) \right] \times \exp \left[ \frac{1}{2} t_2^2 \left( \frac{n-1}{n} \right) \sigma^2 \right] \\ &\quad (\text{On simplification}) \\ &= M(t_1) \cdot M(t_2) \end{aligned}$$

$\Rightarrow$  (a)  $\bar{X}$  and  $X_i - \bar{X}; i = 1, 2, \dots, n$  are independently distributed ... (vi)

and (b)  $\bar{X} \sim N(\mu, \sigma^2/n) \quad \dots(vii)$

and  $X_i - \bar{X} \sim N \left( 0, \frac{n-1}{n} \sigma^2 \right) \quad \dots(viii)$

Since  $\bar{X}$  and  $X_i - \bar{X}$ ;  $i = 1, 2, \dots, n$  are independently distributed,

$$\bar{X} \quad \text{and} \quad s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \quad \dots \text{(viii a)}$$

are independently distributed.

To derive the distribution of  $s^2$ , we note that :

$$\begin{aligned} \sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2, \end{aligned}$$

the product term vanishes since  $\sum_{i=1}^n (X_i - \bar{X}) = 0$ .

$$\therefore \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} + \left[ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right]^2 \quad \dots \text{(ix)}$$

$$\Rightarrow V = W + Z, \quad \dots \text{(ix a)}$$

where  $V = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2$ , being the sum of squares of  $n$  independent standard normal variates is a  $\chi^2_{(n)}$  variate. Hence

$$M_V(t) = (1 - 2t)^{-n/2}; |t| < 1/2, \quad \dots \text{(x)}$$

$$\text{Also } \bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$\therefore Z = \left[ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right]^2 \sim \chi^2_{(1)}$$

$$\Rightarrow M_Z(t) = (1 - 2t)^{-1/2} \quad \dots \text{(xi)}$$

Further, since  $\bar{X}$  and  $s^2$  are independent, [see viii (a)],  $W$  and  $Z$  are independently distributed.

$$\therefore M_V(t) = M_{W+Z}(t) = M_W(t) \cdot M_Z(t)$$

( $\because W$  and  $Z$  are independent).

$$\Rightarrow (1 - 2t)^{-n/2} = M_W(t) \cdot (1 - 2t)^{-1/2} \quad [\text{From (x) and (xi)}]$$

$$\Rightarrow M_W(t) = (1 - 2t)^{-(n-1)/2}, |t| < 1/2$$

which is the m.g.f. of  $\chi^2$ -variante with  $(n - 1)$  d.f. Hence by uniqueness theorem of m.g.f.

$$W = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n-1)} \quad \dots \text{(xii)}$$

**Remarks 1.** p.d.f. of the sample variance  $s^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ .

$$\text{Since } \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 = ns^2/\sigma^2$$

is a  $\chi^2$ -variate with  $(n - 1)$  d.f., we have

$$\begin{aligned} dP(ns^2/\sigma^2) &= \frac{1}{2^{(n-1)/2} \Gamma[(n-1)/2]} \cdot e^{-ns^2/2\sigma^2} \left( \frac{ns^2}{\sigma^2} \right)^{(n-1)/2-1} d(ns^2/\sigma^2) \\ \Rightarrow dP(s^2) &= \frac{(n/2\sigma^2)^{(n-1)/2}}{\Gamma[(n-1)/2]} \cdot e^{-ns^2/2\sigma^2} (s^2)^{(n-3)/2} ds^2, \quad 0 < s^2 < \infty. \end{aligned}$$

2. We have

$$\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\therefore E\left(\frac{ns^2}{\sigma^2}\right) = n - 1$$

$$\Rightarrow \frac{n}{\sigma^2} E(s^2) = (n - 1)$$

$$\Rightarrow E(s^2) = \left( \frac{n-1}{n} \right) \sigma^2 = \left( 1 - \frac{1}{n} \right) \sigma^2 \approx \sigma^2, \text{ for large } n. \quad \dots(*)$$

$$\text{Also } \text{Var}\left(\frac{ns^2}{\sigma^2}\right) = 2(n-1)$$

$$\Rightarrow \frac{n^2}{\sigma^4} \text{Var}(s^2) = 2(n-1)$$

$$\Rightarrow \text{Var}(s^2) = \frac{2}{n} \left( 1 - \frac{1}{n} \right) \sigma^4 \approx \frac{2\sigma^4}{n}, \text{ for large } n. \quad \dots(**)$$

$$\Rightarrow \text{S.E.}(s^2) = \sigma^2 \times \sqrt{2/n} \quad \dots(***)$$

**Theorem 13.6.** Let  $X_i$ , ( $i = 1, 2, \dots, n$ ) be independent  $N(0, 1)$  variates.

Then the conditional distribution of  $\chi^2 = \sum_{i=1}^n X_i^2$ , subject to  $m$  ( $< n$ ) (say), independent homogeneous linear constraints viz.,

$$\left. \begin{array}{l} c_{11}X_1 + c_{12}X_2 + \dots + c_{1n}X_n = 0 \\ c_{21}X_1 + c_{22}X_2 + \dots + c_{2n}X_n = 0 \\ \vdots \qquad \vdots \qquad \vdots \\ c_{m1}X_1 + c_{m2}X_2 + \dots + c_{mn}X_n = 0 \end{array} \right\} \quad \dots(13.12)$$

is also a  $\chi^2$ -distribution with  $(n - m)$  degrees of freedom.

**Proof.** Equivalently, the constraints (13.12) can be expressed as

$$\left. \begin{array}{l} a_{11}X_1 + a_{12}X_2 + \dots + a_{1n}X_n = 0 \\ a_{21}X_1 + a_{22}X_2 + \dots + a_{2n}X_n = 0 \\ \vdots \qquad \vdots \qquad \vdots \\ a_{m1}X_1 + a_{m2}X_2 + \dots + a_{mn}X_n = 0 \end{array} \right\} \quad \dots(13.12a)$$

where  $a_i = (a_{i1}, a_{i2}, \dots, a_{in})$ ;  $i = 1, 2, \dots, m$  are  $m$  unitary, mutually orthogonal vectors.

Let us now transform the variables

$(X_1, X_2, \dots, \bar{X}_m, X_{m+1}, \dots, X_n)$  to  $(Y_1, Y_2, \dots, Y_m, Y_{m+1}, \dots, Y_n)$

by means of a linear orthogonal transformation

$$\mathbf{Y} = \mathbf{AX} \quad \dots(13-12b)$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \\ Y_{m+1} \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \\ a_{m+1,1} & a_{m+1,2} & \dots & a_{m+1,n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \\ X_{m+1} \\ \vdots \\ X_n \end{pmatrix}$$

(13-12b) implies that the constraints (13-12a) are equivalent to

$$Y_i = 0, \quad (i = 1, 2, \dots, m) \quad \dots(13-12c)$$

By Fisher's Lemma (Theorem 13-4)  $Y_i$ , ( $i = 1, 2, \dots, n$ ) are also independent  $N(0, 1)$  variables and

$$\begin{aligned} \sum_{i=1}^n X_i^2 &= \sum_{i=1}^m Y_i^2 && [\because \text{Transformation (13-2b) is orthogonal}] \\ &= \sum_{i=m+1}^n Y_i^2 && [\text{Using (13-12c)}] \end{aligned}$$

Thus the conditional distribution of  $\sum_{i=1}^n X_i^2$  subject to the conditions

(13-12) is same as the unconditional distribution of  $\sum_{i=m+1}^n Y_i^2$ ; where  $Y_i$  ( $i = m+1, m+2, \dots, n$ ) are independent standard normal variates without any constraints on them. Hence

$$\chi^2 = \sum_{i=1}^n X_i^2 = \sum_{i=m+1}^n Y_i^2,$$

being the sum of squares of  $(n - m)$  independent standard normal variates follows  $\chi^2$ -distribution with  $(n - m)$  degrees of freedom.

**Example 13-1.** (a) Show that for 2 d.f. the probability  $P$  of a value of  $\chi^2$  greater than  $\chi_0^2$  is  $\exp(-\frac{1}{2}\chi_0^2)$ , and hence that

$$\chi_0^2 = 2 \log_e (1/P)$$

Deduce the value of  $\chi_0^2$  when  $P = 0.05$ .

(b) Given different probabilities  $P_1, P_2, \dots, P_n$  obtained from  $n$  independent tests of significance, explain how you will pool them to get a single probability in order to decide about the significance of the aggregate of these tests.

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

**Solution.** (a) The p.d.f. of  $\chi^2$ -distribution with 2 d.f. is

$$f(\chi^2) = \left[ \frac{1}{2^{n/2} \Gamma(n/2)} \exp(-\chi^2/2) \cdot (\chi^2)^{(n/2)-1} \right]_{n=2}$$

$$= \frac{1}{2} \exp(-\chi^2/2), 0 \leq \chi^2 < \infty$$

$$\therefore P = P(\chi^2 > \chi_0^2) = \int_{\chi_0^2}^{\infty} \frac{1}{2} \exp(-\chi^2/2) d\chi^2 \quad \dots (*)$$

$$= \frac{1}{2} \left| \exp(-\chi^2/2) \right|_{-\frac{1}{2}}^{\infty} = \exp(-\chi_0^2/2)$$

$$\therefore \log_e P = -\chi_0^2/2$$

$$\Rightarrow \chi_0^2 = -2 \log_e P = 2 \log_e (1/P)$$

When  $P = 0.05$ , we get  $\chi_0^2 = 2 \log_e 20 \approx 3.012$

**Remark.** The value  $\chi_0^2$  of  $\chi^2$  defined in (\*), is known as the significant or critical value [c.f. Remark 4, to Theorem 13-3, page 13-15] of  $\chi^2$  corresponding to the probability level  $P$ . Thus if  $P$  is the significant probability, then

$$\chi^2 = -2 \log_e P = 2 \log_e (1/P) \quad \dots (13.13)$$

is a  $\chi^2$ -variate with 2 d.f.

(b)  $-2 \log_e P_i$  ( $i = 1, 2, \dots, n$ ) are independent  $\chi^2$ -variates each with 2 d.f. (c.f. Remark above and the fact that  $P_i$ 's obtained from independent tests of significance are independent). Hence by additive property of chi-square distribution

$$\chi^2 = \sum_{i=1}^n (-2 \log_e P_i) = 2 \log_e \left( \frac{1}{P_1 P_2 \dots P_n} \right) \quad \dots (13.13a)$$

is a chi-square variate with  $2n$  d.f.

If  $\chi^2 > \chi_{0.05}^2$  for  $2n$  d.f., then we conclude that the pooled result (aggregate of the tests) is significant at 5% level of significance.

**Example 13-2.** (Pearson's  $P_\lambda$ -Statistic). The variables  $X_1, X_2, \dots, X_n$  are independently distributed in the rectangular form

$$dF = dx, 0 \leq x \leq 1$$

Then if  $P = X_1 X_2 \dots X_n$ , show that  $-2 \log_e P$  has  $\chi^2$ -distribution with  $2n$  degrees of freedom. (Aligarh Univ. B.Sc., 1992)

**Solution.**  $-2 \log_e P = -2 \log_e (X_1 X_2 \dots X_n)$

$$= \xi_1 + \xi_2 + \dots + \xi_n = \sum_{i=1}^n \xi_i,$$

where  $\xi_i = -2 \log X_i \Rightarrow X_i = \exp(-\xi_i/2)$ .

The probability function of  $\xi_i$  is given by

$$g(\xi_i) = f(x_i) \left| \frac{dx_i}{d\xi_i} \right|$$

Since  $dF(x) = dx$ ,  $f(x) = 1 \forall x \text{ in } [0, 1]$

$$\therefore g(\xi_i) = 1 \cdot \left| \exp(-\xi_i/2) \times \left( -\frac{1}{2} \right) \right| = \frac{1}{2} \exp(-\xi_i/2)$$

which is the probability function of  $\chi^2$ -distribution with 2 d.f.

$\therefore \xi_i$ , ( $i = 1, 2, \dots, k$ ) are independent  $\chi^2$ -variates each with 2 d.f. Hence by additive property of  $\chi^2$ -distribution,

$$-2 \log_e P = \sum_{i=1}^n \xi_i,$$

is a  $\chi^2$ -variante with  $2n$  d.f.

**Remark.** The significance of this result lies in testing of hypothesis as explained in Example 13.1.

**Example 13.3.** Show that if  $v$  is even,

$$\begin{aligned} P &= \frac{1}{2^{(v-2)/2} \Gamma(v/2)} \int_{-\infty}^{\infty} \exp(-\chi^2/2) \chi^{v-1} d\chi \\ &= \exp(-\chi^2/2) \left[ 1 + (\chi^2/2) + \frac{\chi^4}{2 \cdot 4} + \dots + \frac{\chi^{v-2}}{2 \cdot 4 \dots (v-2)} \right] \end{aligned}$$

and hence the values of  $P$  for a given  $\chi^2$  can be derived from tables of Poisson's exponential limit.

**Solution.** Let us consider the incomplete Gamma-integral

$$I_r = \frac{1}{r!} \int_{\beta}^{\infty} e^{-t} t^r dt,$$

where  $r$  is a positive integer. Integrating by parts, we get

$$I_r = \left[ -\frac{e^{-t} t^r}{r!} \right]_{\beta}^{\infty} + \frac{1}{(r-1)!} \int_{\beta}^{\infty} e^{-t} t^{r-1} dt = \frac{e^{-\beta} \beta^r}{r!} + I_{r-1}$$

which is a reduction formula. Repeated application of this gives

$$I_r = \frac{e^{-\beta} \beta^r}{r!} + \frac{e^{-\beta} \beta^{r-1}}{(r-1)!} + \dots + \frac{e^{-\beta} \beta^2}{2!} + \frac{e^{-\beta} \beta}{1!} + I_0$$

$$\text{But } I_0 = \int_{\beta}^{\infty} e^{-t} dt = \left[ -e^{-t} \right]_{\beta}^{\infty} = e^{-\beta}$$

$$\therefore \frac{1}{r!} \int_{\beta}^{\infty} e^{-t} t^r dt = e^{-\beta} \left[ 1 + \beta + \frac{\beta^2}{2!} + \dots + \frac{\beta^{r-1}}{(r-1)!} + \frac{\beta^r}{r!} \right].$$

Putting  $\beta = \chi^2/2$  and  $r = \frac{v-2}{2} = \frac{v}{2} - 1$ , (since  $r$  is an integer,  $v=2r+2$  must be even), we get

$$\begin{aligned} & \frac{1}{\{(v/2) - 1\}!} \int_{\chi^2/2}^{\infty} e^{-t} t^{(v/2)-1} dt \\ &= \exp(-\chi^2/2) \left[ 1 + \frac{\chi^2}{2} + \frac{\chi^4}{2.4} + \frac{\chi^6}{2.4.6} + \dots + \frac{\chi^{v-2}}{2.4.6\dots(v-2)} \right] \quad \dots(*) \end{aligned}$$

Taking  $t = \chi^2/2$  in the integral on the L.H.S., we get

$$\begin{aligned} \text{L.H.S.} &= \frac{1}{\{(v/2) - 1\}!} \int_{\chi}^{\infty} \exp(-\chi^2/2) (\chi^2/2)^{(v/2)-1} d(\chi^2/2) \\ &= \frac{1}{2^{(v-2)/2} \Gamma(v/2)} \int_{\chi}^{\infty} \exp(-\chi^2/2) \chi^{v-1} d\chi. \quad \dots(**) \end{aligned}$$

From (\*) and (\*\*), we get the required result.

Let the given value of  $\chi^2$  be  $\chi_0^2$ , then

$$\begin{aligned} P = P(\chi^2 > \chi_0^2) &= \frac{1}{2^{(v-2)/2} \Gamma(v/2)} \int_{\chi_0}^{\infty} \exp(-\chi^2/2) \chi^{v-1} d\chi \\ &= \exp(-\chi_0^2/2) \left[ 1 + \frac{\chi_0^2}{2} + \frac{\chi_0^4}{2.4} + \dots + \frac{\chi_0^{v-2}}{2.4\dots(v-2)} \right] \\ &= e^{-\lambda} \left[ 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots + \frac{\lambda^{v-1}}{[(v/2) - 1]!} \right], \end{aligned}$$

where  $\lambda = \chi_0^2/2$ .

The terms on the right hand side viz.,  $e^{-\lambda}$ ,  $\lambda e^{-\lambda}$ ,  $\frac{\lambda^2}{2!} e^{-\lambda}$ ...etc. are the successive terms of the Poisson distribution with parameter  $\lambda = \chi_0^2/2$ .

Hence the result.

**Example 13-4.** If  $X$  and  $Y$  are independent normal variates with means  $\mu_1, \mu_2$  and variances  $\sigma_1^2, \sigma_2^2$  respectively, derive the distribution of

$$Z = (X - \mu_1)/(Y - \mu_2).$$

What is the name of the distribution so obtained? Mention one important property of this distribution.

**Solution.** Here  $Z^2 = \frac{(X - \mu_1)^2}{(Y - \mu_2)^2} \Rightarrow \frac{\sigma_2^2}{\sigma_1^2} \cdot Z^2 = \frac{[(X - \mu_1)/\sigma_1]^2}{[(Y - \mu_2)/\sigma_2]^2}$

But  $(X - \mu_1)/\sigma_1$  and  $(Y - \mu_2)/\sigma_2$ , being the squares of independent standard normal variates, are independent  $\chi^2$ -variates with 1 d.f. each.

Thus  $\frac{\sigma_2^2 z^2}{\sigma_1^2}$ , being the quotient of two independent  $\chi^2$ -variates each with 1 d.f. is a  $\beta_2(\frac{1}{2}, \frac{1}{2})$  variate (c.f. Theorem 13-1).

Hence its probability differential is given by

$$dF\left(\frac{\sigma_2^2 z^2}{\sigma_1^2}\right) = \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \times \frac{(\sigma_2^2 z^2 / \sigma_1^2)^{(1/2)-1}}{\left(1 + \frac{\sigma_2^2 z^2}{\sigma_1^2}\right)^{\frac{1}{2} + \frac{1}{2}}} d(\sigma_2^2 z^2 / \sigma_1^2)$$

$$= \frac{\Gamma(\frac{1}{2} + \frac{1}{2})}{\Gamma(\frac{1}{2}) \Gamma(\frac{1}{2})} \cdot \frac{\left(\frac{\sigma_2^2 z^2}{\sigma_1^2}\right)^{\frac{1}{2}-1}}{\left(1 + \frac{\sigma_2^2 z^2}{\sigma_1^2}\right)} \cdot \frac{\sigma_2^2}{\sigma_1^2} dz^2$$

$$= \frac{\sigma_1 \sigma_2}{\pi(\sigma_1^2 + \sigma_2^2 z^2)} dz^2, \quad 0 < z^2 < \infty \quad [\because \Gamma(1/2) = \sqrt{\pi}]$$

Thus the probability differential of Z is given by

$$dF(z) = f(z)dz = \frac{\sigma_1 \sigma_2}{\pi(\sigma_1^2 + \sigma_2^2 z^2)} dz, \quad -\infty < z < \infty$$

If  $\sigma_1 = \sigma_2 = 1$ , then it conforms to standard Cauchy distribution,

$$dF(z) = \frac{1}{\pi} \cdot \frac{dz}{(1 + z^2)}, \quad -\infty < z < \infty$$

[For its properties c.f. Chapter 8].

$$\text{Aliter} \quad Z = \frac{X - \mu_1}{Y - \mu_2},$$

$$\Rightarrow \frac{\sigma_2}{\sigma_1} Z = \frac{(X - \mu_1)/\sigma_1}{(Y - \mu_2)/\sigma_2}$$

Now  $\frac{\sigma_2}{\sigma_1} Z$ , being the ratio of two independent standard normal variates is a standard Cauchy variate

$$\therefore dF\left(\frac{\sigma_2}{\sigma_1} z\right) = \frac{d\left(\frac{\sigma_2}{\sigma_1} z\right)}{\pi \left[ 1 + \left( \frac{\sigma_2}{\sigma_1} z \right)^2 \right]} \\ = \frac{\sigma_1 \sigma_2}{\pi(\sigma_1^2 + \sigma_2^2 z^2)} dz, \quad -\infty < z < \infty$$

**Example 13-5.**  $X_i, (i = 1, 2, \dots, n)$  are independently and normally distributed with zero mean and common variance  $\sigma^2$ .

Let  $\xi_i = \sum_{j=1}^n c_{ij} X_j, ; i = 1, 2, \dots, n$ , where  $\sum_{j=1}^n c_{ij} c_{i'j} = \delta_{ii}'$

where  $\delta_{ii}'$  is Kronecker delta. Show that

$$\left[ \sum_{i=1}^n X_i^2 - \sum_{i=1}^p \xi_i^2 \right] \not\perp \sigma^2$$

is distributed as  $\chi^2$ -variate with  $(n-p)$  degrees of freedom.

[Delhi Univ. M.Sc. (Stat.), 1990]

**Solution.** Since  $\delta_{ii}' = \sum_{j=1}^n c_{ij} c_{ij}'$

is a Kronecker delta, we have

$$\sum_{j=1}^n c_{ij} c_{ij}' = \begin{cases} 0, & i \neq i' \\ 1, & i = i' \end{cases}$$

i.e.,  $X_i$ 's are transformed to  $\xi_i$ 's by means of a linear orthogonal transformation. Hence by Fisher's Lemma,  $\xi_i$ , ( $i = 1, 2, \dots, n$ ) are independent normal variates with zero mean and common variance  $\sigma^2$ .

Since the transformation is orthogonal, we have

$$\sum_{i=1}^n X_i^2 = \sum_{i=1}^n \xi_i^2$$

$$\therefore \frac{\left[ \sum_{i=1}^n X_i^2 - \sum_{i=1}^p \xi_i^2 \right]}{\sigma^2} = \frac{\sum_{i=1}^n \xi_i^2 - \sum_{i=1}^p \xi_i^2}{\sigma^2} = \frac{\sum_{i=p+1}^n \xi_i^2}{\sigma^2}$$

Now  $\frac{\sum_{i=p+1}^n \xi_i^2}{\sigma^2} = \sum_{i=p+1}^n (\xi_i/\sigma)^2$ ,

being the sum of the squares of  $(n-p)$  independent standard normal variates is a  $\chi^2$ -variate with  $(n-p)$  degrees of freedom. Hence the result.

**Example 13-6.** Show that the m.g.f. of  $Y = \log \chi^2$ , where  $\chi^2$  follows chi-square distribution with  $n$  d.f., is given by

$$M_Y(t) = 2^t \Gamma\left(\frac{n}{2} + t\right) \not\perp \Gamma(n/2)$$

If  $\chi_1^2$  and  $\chi_2^2$  are independent  $\chi^2$ -variates each with  $n$  d.f. and  $U = \chi_1^2/\chi_2^2$ , deduce that for positive integer  $k$ ,

$$E(U^k) = \Gamma\left(\frac{n}{2} + k\right) \Gamma\left(\frac{n}{2} - k\right) \not\perp \left[ \Gamma\left(\frac{n}{2}\right) \right]^2$$

**Solution.**  $y = \log \chi^2 \Rightarrow \chi^2 = e^y \Rightarrow d\chi^2 = e^y dy$ .

The probability differential of  $\chi^2$  viz.,

$$dP(\chi^2) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-\chi^2/2} (\chi^2)^{\frac{n}{2}-1} d\chi^2, 0 < \chi^2 < \infty$$

transforms to

$$\begin{aligned}
 dG(y) &= \frac{1}{2^{n/2} \Gamma(n/2)} \exp \left[ -\frac{1}{2} e^y + \frac{ny}{2} \right] dy, \quad -\infty < y < \infty \\
 M_Y(t) &= \frac{1}{2^{n/2} \Gamma(n/2)} \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} e^y + \frac{ny}{2} + ty \right] dy \\
 &= \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^{\infty} e^{-z} (2z)^{\frac{n}{2}+t} \frac{dz}{z}, \quad (2z = e^y) \\
 &= \frac{2^t}{\Gamma(n/2)} \int_0^{\infty} e^{-z} z^{\frac{n}{2}+t-1} dz \\
 &= 2^t \Gamma\left(\frac{n}{2} + t\right) / \Gamma\left(\frac{n}{2}\right) \quad \dots(*) \\
 E(U^k) &= E\left[\left(\frac{\chi_1^2}{\chi_2^2}\right)^k\right] = E\left[\exp\left\{\log\left(\frac{\chi_1^2}{\chi_2^2}\right)^k\right\}\right] \\
 &= E\left[e^{k \log \chi_1^2 - k \log \chi_2^2}\right] \\
 &= E\left(e^{k \log \chi_1^2}\right) \cdot E\left(e^{-k \log \chi_2^2}\right) \\
 &\quad [\because \chi_1^2 \text{ and } \chi_2^2 \text{ are independent}] \\
 &= M_{\log \chi_1^2}(k) \cdot M_{\log \chi_2^2}(-k) \\
 &= \frac{2^k \Gamma\left(\frac{n}{2} + k\right)}{\Gamma(n/2)} \cdot \frac{2^{-k} \Gamma\left(\frac{n}{2} - k\right)}{\Gamma(n/2)} \quad [\text{From (*)}] \\
 &= \Gamma\left(\frac{n}{2} + k\right) \Gamma\left(\frac{n}{2} - k\right) / \left[\Gamma\left(\frac{n}{2}\right)\right]^2
 \end{aligned}$$

**Example 13-7.** If  $X$  is chi-square variate with  $n.d.f.$ , then prove that for large  $n$ ,  $\sqrt{2}X \sim N(\sqrt{2n}, 1)$  [Delhi Univ. B.Sc. (Stat. Hons.), 1989]

**Solution.** We have  $E(X) = n$ ,  $\text{Var}(X) = 2n$

$$Z = \frac{X - E(X)}{\sigma_X} = \frac{X - n}{\sqrt{2n}} \sim N(0, 1), \text{ for large } n.$$

Consider,

$$\begin{aligned}
 P\left(\frac{X - n}{\sqrt{2n}} \leq z\right) &= P(X \leq n + z\sqrt{2n}) \\
 &= P[\sqrt{2}X \leq (2n + 2z\sqrt{2n})^{1/2}] \\
 &= P\left[\sqrt{2}X \leq \sqrt{2n} \left(1 + z\sqrt{\frac{2}{n}}\right)^{1/2}\right] \\
 &= P\left[\sqrt{2}X \leq \sqrt{2n} \left(1 + \frac{z}{\sqrt{2n}} - \frac{z^2}{4n} + \dots\right)\right].
 \end{aligned}$$

$$\begin{aligned} & \approx P[\sqrt{2X} \leq \sqrt{2n} + z], \text{ for large } n. \\ & = P[\sqrt{2X} - \sqrt{2n} \leq z], \text{ for large } n. \end{aligned}$$

Since for large  $n$ ,  $(X - n)/\sqrt{2n} \sim N(0, 1)$ , we conclude that

$$\begin{aligned} & \sqrt{2X} - \sqrt{2n} \sim N(0, 1) \text{ for large } n. \\ \Rightarrow & \sqrt{2X} \text{ is asymptotically } N(\sqrt{2n}, 1). \end{aligned}$$

**Remark.** This approximation is often used for the value of  $n$  larger than 30. This result does not reflect anything as to how good the approximation is, for moderate values of  $n$ . R.A. Fisher has proved that the approximation is improved by taking  $\sqrt{(2n - 1)}$  instead of  $\sqrt{2n}$ . A still better approximation is  $(\chi^2/n)^{1/2} \sim N\left(1 - \frac{2}{9n}, \frac{2}{9n}\right)$ .

**Example 13.8.** For a chi-square distribution with  $n$  d.f. establish the following recurrence relation between the moments :

$$\mu_{r+1} = 2r(\mu_r + n\mu_{r-1}), r \geq 1.$$

Hence find  $\beta_1$  and  $\beta_2$ .

[Delhi Univ.B.Sc. (Stat. Hons.), 1991]

**Solution.** If  $X \sim \chi^2_{(n)}$  then its m.g.f. about origin is

$$M_X(t) = E(e^{tX}) = (1 - 2t)^{-n/2}; t < \frac{1}{2} \quad \dots(*)$$

Also  $E(X) = n = \mu$  (say).

Hence m.g.f. about mean, say,  $M(t)$  is

$$\begin{aligned} M(t) &= M_{X-\mu}(t) = E(e^{t(X-\mu)}) = e^{-\mu t} \cdot E(e^{tX}) \\ &= e^{-\mu t} (1 - 2t)^{-n/2} \end{aligned}$$

Taking logarithms of both sides, we get

$$\log M(t) = -\mu t - \frac{n}{2} \log(1 - 2t) \quad [\text{Using (*)}]$$

Differentiating w.r. to  $t$ , we have

$$\begin{aligned} \frac{M'(t)}{M(t)} &= -n + \frac{n}{2} \cdot \frac{2}{(1 - 2t)} = \frac{2nt}{(1 - 2t)} \\ \Rightarrow (1 - 2t) M'(t) &= 2nt M(t) \end{aligned}$$

Differentiating  $r$  times w.r. to  $t$  by Leibnitz theorem, we get

$$(1 - 2t) M^{r+1}(t) + r(-2) M^r(t) = 2nt M^r(t) + 2nr M^{r-1}(t)$$

Putting  $t = 0$  and using the relation,

$$\mu_r = \left[ \frac{d^r}{dt^r} M(t) \right]_{t=0} = M'(0), \text{ we get}$$

$$\begin{aligned} \mu_{r+1} - 2r \mu_r &= 2nr \mu_{r-1} \\ \Rightarrow \mu_{r+1} &= 2r(\mu_r + n\mu_{r-1}), r \geq 1. \end{aligned}$$

Substituting  $r = 1, 2, 3$ ; we get

$$\begin{aligned}\mu_2 &= 2n\mu_0 = 2n \\ \mu_3 &= 4(\mu_2 + n\mu_1) = 8n \quad [\because \mu_1 = 0 \text{ and } \mu_0 = 1] \\ \mu_4 &= 6(\mu_3 + n\mu_2) = 48n + 12n^2 \\ \therefore \beta_1 &= \frac{\mu_3^2}{\mu_2^2} = \frac{8}{n} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{12}{n}\end{aligned}$$

**EXERCISE 13(a)**

1. (a) Derive the p.d.f. of chi-square distribution with  $n$  degrees of freedom.  
 (b) If  $X$  has a chi-square distribution with  $n$  d.f., find m.g.f.  $M_X(t)$ .

Deduce that :

- (i)  $\mu_r' = EX^r = 2^r \Gamma[(n/2) + r] / \Gamma(n/2)$ .
- (ii)  $k_r = r$ th cumulant  $= n 2^{r-1} (r-1)!$
- (iii)  $k_1 k_3 = 2k_2^2, 2\beta_2 - 3\beta_1 - 6 = 0$ .

2. If  $X \sim \chi^2_{(n)}$ , show that :

- (i) Mode is at  $x = n - 2$ .
- (ii) The points of inflexion are equidistant from the mode.

**Hint.** Points of inflexion are at  $x = (n-2) \pm [2(n-2)]^{1/2}$

3. If  $X \sim \chi^2_{(n)}$ , obtain the m.g.f. of  $X$ . Hence find the m.g.f. of standard chi-square variate and obtain its limiting form as  $n \rightarrow \infty$ . Also interpret the result.

4. (a) Let  $X \sim N(0, 1)$  and  $Y = X^2$ . Calculate  $E(Y)$  in two different ways.

**Ans.**  $E(Y) = 1$ . (Use Normal distribution and chi-square distribution).

- (b) Let  $X_1$  and  $X_2$  be independent standard normal variates and let

$$Y = (X_2 - X_1)^2/2. \text{ Find the distribution of } Y.$$

**Ans.**  $Y \sim \chi^2_{(1)}$ .

5. If  $X_1, X_2, \dots, X_n$  are i.i.d. exponential variates with parameter  $\lambda$ , prove that

$$2\lambda \sum_{i=1}^n X_i \sim \chi^2_{(2n)}$$

6. (a) If  $X \sim U[0, 1]$ , show that  $-2 \log X \sim \chi^2_{(2)}$ .

Hence show that if  $X_1, X_2, \dots, X_n$  are i.i.d.  $U[0, 1]$  variates, and if

$$P = X_1 X_2 \dots X_n, \text{ then } -2 \log P \sim \chi^2_{(2n)}$$

**Hint.** Find m.g.f. of  $-2 \log X$ .

- (b) If  $X_1, X_2, \dots, X_n$  are independent random variables with continuous distribution functions  $F_1, F_2, \dots, F_n$  respectively, show that

$$-2 \log [F_1(X_1) \cdot F_2(X_2) \dots F_n(X_n)] \sim \chi^2_{(2n)}$$

**Hint.** Use  $F(X) \sim U[0, 1]$  and Part (a) above.

7. (a) Let  $X$  and  $Y$  be two independent random variables having chi-square distribution with degrees of freedom  $m$  and  $n$  respectively.

(i) Obtain the distribution of  $U = \frac{X}{X + Y}$ .

(ii) When  $m = n$ , show that the distribution of  $U$  is symmetrical about  $\frac{1}{2}$ .

Hence or otherwise derive the  $r$ th moment about the mean of  $U$  when  $m = n$ .

(iii) Deduce the distribution of  $U$  when  $m = n = 1$ .

(b) If  $X$  and  $Y$  are independently distributed chi-square variates with  $m$  and  $n$  degrees of freedom respectively, show that  $U = X + Y$  and  $V = X/Y$  are independently distributed. [Gujarat Univ. B.Sc., 1992]

(c) If  $\chi_1^2$  and  $\chi_2^2$  are independent  $\chi^2$  variates with  $n_1$  and  $n_2$  degrees of freedom respectively, then show that :

(i)  $\chi^2 = \chi_1^2 + \chi_2^2$  is a  $\chi^2$ -variate with  $(n_1 + n_2)$  degrees of freedom

(ii)  $T^2 = \frac{\chi_1^2}{\chi_2^2}$  is a  $\beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$  variate.

[Delhi Univ. B.Sc. (Maths. Hons.), 1987]

8. If  $X_i$  ( $i = 1, 2, \dots, n$ ) are  $n$  independent normal variates with zero means and unit variances, show that  $\sum_{i=1}^n X_i$  and  $\sum_{i=1}^n (X_i - \bar{X})^2$  are independently

distributed.

Hence or otherwise obtain the distribution of

$$U = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

9. (a) Prove that  $\frac{nS^2}{\sigma^2}$  is distributed like  $\chi^2$  with  $(n - 1)$  degrees of freedom, where  $S^2$  and  $\sigma^2$  are the variances of sample (of size  $n$ ) and the population respectively. [Burdwan Univ. B.Sc. (Maths.) Hons., 1992]

(b) Let  $X/\sigma_1^2$  and  $Y/\sigma_2^2$  be two independent chi-square variates with  $n$  and  $m$  degrees of freedom respectively. Find an unbiased estimate of  $(\sigma_1/\sigma_2)^2$  and find its variance. Show that  $X/Y$  and  $(X/\sigma_1^2) + (Y/\sigma_2^2)$  are independently distributed. Name the distributions of  $X/Y$  and  $(X/\sigma_1^2) + (Y/\sigma_2^2)$ .

10.  $X$  denotes the random variable with chi-square distribution having  $n$  degrees of freedom. Show that for suitably chosen constants  $a_n$  and  $b_n$ , the moment generating function of  $\frac{X - a_n}{b_n}$  tends to that of the standard normal distribution as  $n \rightarrow \infty$ . From this what would you conclude about the behaviour, for large  $n$ , of  $P\left(\frac{X - a_n}{b_n} \leq x\right)$ ?

11. Show that

$$P\{\chi^2_{2v+2} \geq 2\lambda\} = \frac{1}{v!} \int_{\lambda}^{\infty} e^{-y} y^v dy = \sum_{r=0}^{v} \frac{e^{-\lambda} \lambda^r}{r!}$$

where  $\lambda = \frac{1}{2} \chi_0^2$ .

Explain the uses of this result. [Delhi Univ. B.Sc. (Stat. Hons.), 1990]

12.  $X$  is a Poisson variate with parameter  $\lambda$  and  $\chi^2$  is a chi-square variate with  $2k$  d.f. Prove that for all positive integers  $k$ ,

$$P\{X \leq k - 1\} = P\{\chi^2 > 2\lambda\}$$

$$\begin{aligned} \text{Hint. } P(\chi^2 > 2\lambda) &= \frac{1}{2^k \Gamma(k)} \int_{2\lambda}^{\infty} \exp(-\frac{1}{2}\chi^2) (\chi^2)^{k-1} d\chi^2 \\ &= \frac{1}{(k-1)!} \int_{\lambda}^{\infty} e^{-y} y^{k-1} dy \\ &= \frac{1}{(k-1)!} \left\{ \lambda^{k-1} e^{-\lambda} + (k-1) \int_{\lambda}^{\infty} e^{-y} y^{k-2} dy \right\} \end{aligned}$$

By repeated integration, we get the required result.

13. Let  $X_1, X_2, \dots, X_m$  and  $Y_1, Y_2, \dots, Y_n$  be independent random samples from a normal population with mean zero and variance  $\sigma^2$ . Let their means be  $\bar{X}$  and  $\bar{Y}$  and their variances be  $S_X^2$  and  $S_Y^2$  respectively.

Let the pooled variance  $S_p^2$  be defined as :

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{(m+n-2)}$$

Prove that  $(\bar{X} - \bar{Y})$  and  $(m+n-2) S_p^2 / \sigma^2$  are independently distributed, the former as a normal variate with zero mean and variance  $\sigma^2 \{ (1/m) + (1/n) \}$  and the latter as a chi-square variate with  $(m+n-2)$  d.f.

[Nagpur Univ B.E., 1992]

14. If  $X$  is a random variable following Poisson distribution with parameter  $\lambda$ , and  $\lambda$  is also a random variable so that  $2\alpha\lambda$  is a chi-square variate with  $2p$  degrees of freedom, obtain the unconditional distribution of  $X$ . Give the name of this distribution and find its mean.

15.  $X_1, X_2$ , and  $X_3$  denote independent central chi-square variates with  $v_1, v_2$  and  $v_3$  d.f. respectively.

(i) Show that  $X_1/(X_1 + X_2)$  is independently distributed of

$$(X_1 + X_2)/(X_1 + X_2 + X_3).$$

(ii) Obtain the joint density function of the distribution of

$$X = X_1/(X_1 + X_2 + X_3) \text{ and } Y = X_2/(X_1 + X_2 + X_3)$$

(iii) Hence or otherwise obtain the mean and variance of  $X$  and  $Y$  and  $\text{Cov}(X, Y)$ .

16. Prove that each linear constraint on  $\{f_i\}$ ,  $i = 1, 2, \dots, n$  reduces by unity the number of degrees of freedom of the chi-square,

$$\sum_{i=1}^n \{(f_i - e_i)^2/e_i\},$$

where  $e_i = E(f_i)$ .

17. If  $X$  follows a chi-square distribution with  $n$  d.f. so that  $E(X) = n$  and  $V(X) = 2n$ , prove that  $(X - n)/\sqrt{2n}$  is a  $N(0, 1)$ , for large  $n$ .

18. If  $ydx$  is the probability that  $X$  lies between  $x$  and  $x + dx$  and if  $y$  is given by the solution of the differential equation

$$\frac{dy}{dx} = \frac{y(a-x)}{bx+c}$$

show that, (for suitable values of the constants  $a$ ,  $b$  and  $c$ ), a certain linear function of  $X$  has the  $\chi^2$ -distribution with  $n$  degrees of freedom, where

$$n = 2 \left( 1 + \frac{a}{b} + \frac{c}{b^2} \right)$$

19. If  $X_1$  and  $X_2$  are independently distributed, each as  $\chi^2$  variate with 2 d.f., show that the density function of  $Y = \frac{1}{2}(X_1 - X_2)$  is

$$g(y) = \frac{1}{2} e^{-|y|}, -\infty < y < \infty.$$

20. If  $X$  and  $Y$  are independent r.v.'s having rectangular distribution in the interval  $(0, 1)$ , show that

$$U = \sqrt{-2 \log X} \cos 2\pi Y \text{ and } V = \sqrt{-2 \log X} \sin 2\pi Y$$

are independently distributed as  $N(0, 1)$ . Hence show that  $U^2$  and  $V^2$  are independently distributed as  $\chi^2$ -variates, each with 1 d.f.

$$\text{Hint. } \frac{1}{J} = \frac{\partial(u, v)}{\partial(x, y)} = -\frac{2\pi}{x}$$

$$\text{and } u^2 + v^2 = -2 \log x \Rightarrow x = \exp \left[ -\frac{1}{2}(u^2 + v^2) \right]$$

21. Find the p.d.f. of  $\chi_n = +\sqrt{\chi_n^2}$ , where  $\chi_n^2$  is a  $\chi^2$ -variate with  $n$  d.f. and show that

$$\mu_r' = E(\chi_n^r) = 2^{r/2} \frac{\Gamma[(n+r)/2]}{\Gamma(n/2)}$$

Hence establish that for large  $n$ ,

$$E(\chi_n^2) = [E(\chi_n)]^2$$

$$[\text{Hint. } E(\chi_n^2) = n \text{ and } E(\chi_n) = \mu_1' = 2^{1/2} \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)}]$$

Now use  $\frac{\Gamma(n+k)}{\Gamma n} \approx n^k$ , for large values of  $n$ . [c.f. Remark to § 14.5.7]

22. Let

$$P_x = \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^x w^{(n-2)/2} e^{-w/2} dw, \quad x > 0.$$

Show that

$$x < \frac{n}{1 - P_x}$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

23. Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$ , and  $k$  be a positive integer. Find  $E(S^{2k})$ . In particular, find  $E(S^2)$  and  $\text{Var}(S^2)$ .

$$\left[ S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2 ; \bar{X} = n^{-1} \sum_{i=1}^n X_i \right]$$

$$\text{Ans. } E(S^{2k}) = \left( \frac{2\sigma^2}{n-1} \right)^k \cdot \frac{\Gamma\left(k + \frac{n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} ; k > 0, n > 1$$

$$E(S^2) = \sigma^2, \text{ Var}(S^2) = 2\sigma^4/(n-1).$$

24. Let  $X_1$  and  $X_2$  be independent random variables, each  $N(0, 1)$ . Find the joint distribution of  $Y_1 = X_1^2 + X_2^2$  and  $Y_2 = X_1/X_2$ . Find the marginal distributions of  $Y_1$  and  $Y_2$ . Are  $Y_1$  and  $Y_2$  independent?

**Ans.**  $Y_1 \sim \chi^2_{(2)}$  and  $Y_2$  is standard Cauchy variate. Yes.

25. Let  $X_1$  and  $X_2$  be independent standard normal variates. Let

$$Y_1 = X_1 + X_2 \text{ and } Y_2 = X_1^2 + X_2^2.$$

(i) Show that the joint m.g.f. of  $Y_1$  and  $Y_2$  is :

$$M(t_1, t_2) = \frac{1}{1 - 2t_2} \exp \left[ \frac{t_1^2}{1 - 2t_2} \right] ; -\infty < t_1 < \infty, -\infty < t_2 < \frac{1}{2}$$

(ii) Hence or otherwise, show that

$Y_1$  is a normal variate and  $Y_2$  is a chi-square variate.

(iii) Are  $Y_1$  and  $Y_2$  independent? If not, find the correlation coefficient of  $Y_1$  and  $Y_2$ .

[Delhi Univ. B.Sc. (Maths. Hons.), 1989]

**Ans.**  $(Y_1, Y_2)$  are not independent.  $\rho(Y_1, Y_2) = 0$ .

26. If  $X_1, X_2, \dots, X_n$  are independently and normally distributed with the same mean but different variances  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$  and assuming that

$$U = \frac{\sum_i (X_i / \sigma_i^2)}{\sum_i (1/\sigma_i^2)} \text{ and } V = \sum_{i=1}^n \left[ \frac{(X_i - U)^2}{\sigma_i^2} \right]$$

are independently distributed, show that  $U \sim N\{0, 1/(\sum_i \sigma_i^2)\}$  and  $V$  has  $\chi^2$  distribution with  $(n-1)$  d.f.

27. If  $X_1, X_2, \dots, X_n$  is a random sample from  $N(\mu, \sigma^2)$ , find the mean and variance of

$$S = \left[ \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1) \right]^{1/2}$$

[Delhi Univ. B.Sc. (Maths. Hons.), 1988]

28. Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(0, 1)$ . Define :

$$\bar{X}_k = \frac{1}{k} \sum_{i=1}^k X_i \text{ and } \bar{X}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^n X_i$$

- (a) What is the distribution of  $\frac{1}{2}(\bar{X}_k + \bar{X}_{n-k})$ ?
- (b) What is the distribution of  $k\bar{X}_k^2 + (n-k)\bar{X}_{n-k}^2$ ?
- (c) What is the distribution of  $X_i^2/X_j^2$ ,  $i \neq j$ ?
- (d) What is the distribution of  $X_i/X_j$ ,  $i \neq j$ ?

**Ans.** (a)  $N\left(0, \frac{n}{4k(n-k)}\right)$ ; (b)  $\chi^2_{(2)}$

(c)  $\beta_2\left(\frac{1}{2}, \frac{1}{2}\right)$  or  $F(1, 1)$  [See § 14.5]

(d) Standard Cauchy distribution.

### OBJECTIVE TYPE QUESTIONS

I. Choose the correct answer from B and match it with each item in A.

A

(a) $\beta_2$ for a chi-square distribution	(1) $(1-2it)^{-n/2}$
(b) $\beta_1$ for a chi-square distribution	(2) $8/n$
(c) Mean for a chi-square distribution	(3) $2n$
(d) Variance for a chi-square distribution	(4) $(1-2t)^{-n/2}$
(e) Characteristic function for $\chi^2$ distribution	(5) $(12/n) + 3$
(f) Mode of $\chi^2$ -distribution	(6) $\sqrt{2/n}$
(g) M.G.F. of $\chi^2$ -distribution	(7) $n$
(h) Skewness of $\chi^2$ -distribution	(8) $(n-2)$

B

II. State which of the following statement are True and which are False. In case of false statements, give the correct statement.

- (i) Normal distribution is particular case of  $\chi^2$ -distribution for one d.f.
- (ii) For large degrees of freedom, chi-square distribution tends to normal distribution.
- (iii) The sum of independent chi-square variates is also a chi-square variate.
- (iv) For the validity of  $\chi^2$ -test, it is always necessary that the sample observations should be independent.
- (v) The chi-square distribution maintains its character of continuity if cell frequency is less than 5.

- (vi) Each linear constraint reduces the number of degrees of freedom of chi-square by unity.
- (vii) In a chi-square test of goodness of fit, if the calculated value of  $\chi^2$  is zero then the fit is a bad fit.

**III. Mention the correct answer :**

- (i) The mean of a chi-square distribution with  $n$  d.f. is
  - (a)  $2n$ , (b)  $n^2$ , (c)  $\sqrt{n}$ , (d)  $n$
- (ii) The characteristic function of chi-square distribution is
  - (a)  $(1 - 2it)^{n/2}$ , (b)  $(1 + 2it)^{n/2}$ , (c)  $(1 - 2it)^{-n/2}$
- (iii) The range of  $\chi^2$ -variate is
  - (a)  $-\infty$  to  $+\infty$ , (b) 0 to  $\infty$ , (c) 0 to 1, (d)  $-\infty$  to 0.
- (iv) The skewness in a chi-square distribution will be zero if
  - (a)  $n \rightarrow \infty$ , (b)  $n = 0$ , (c)  $n = 1$ , (d)  $n < 0$
- (v) The moment generating function of a  $\chi^2$ -distribution with  $n$  degrees of freedom is
  - (a)  $(1 - t)^{-n/2}$ , (b)  $(1 - 2t)^{-n/2}$ , (c)  $(1 - 3t)^{-n/2}$ , (d)  $(1 - 2t)^{n/2}$
- (iv) Chi-square distribution is
  - (a) Continuous, (b) multimodal, (c) symmetrical.

**IV. Mention some prominent features of the chi-square distribution with  $n$  degrees of freedom.**

**V.** If  $X$  and  $Y$  are independent random variables having chi-square distribution with  $m$  and  $n$  degrees of freedom respectively, write down the distributions of (i)  $X + Y$ , (ii)  $X/Y$ , (iii)  $X/(X + Y)$ .

- VI.** (a) For how many degrees of freedom does the  $\chi^2$ -distribution reduce to negative exponential distribution ?  
 (b) Give an example of two independent variates none of which is a chi-square variate, although their sum is a chi-square variate.

**13-7. Applications of Chi-square Distribution.**  $\chi^2$ -distribution has a large number of applications in Statistics, some of which are enumerated below :

- (i) To test if the hypothetical value of the population variance is  $\sigma^2 = \sigma_0^2$  (say).
- (ii) To test the 'goodness of fit'.
- (iii) To test the independence of attributes.
- (iv) To test the homogeneity of independent estimates of the population variance.
- (v) To combine various probabilities obtained from independent experiments to give a single test of significance.
- (vi) To test the homogeneity of independent estimates of the population correlation coefficient.

In the following sections we shall briefly discuss these applications.

**13.7.1. Chi-square Test for Population Variance.** Suppose we want to test if a random sample  $x_i$ , ( $i = 1, 2, \dots, n$ ) has been drawn from a normal population with a specified variance  $\sigma^2 = \sigma_0^2$ , (say).

Under the null hypothesis that the population variance is  $\sigma^2 = \sigma_0^2$ , the statistic

$$\chi^2 = \sum_{i=1}^n \left[ \frac{(x_i - \bar{x})^2}{\sigma_0^2} \right] = \frac{1}{\sigma_0^2} \left[ \sum_{i=1}^n x_i^2 - \frac{(\sum x_i)^2}{n} \right] = ns^2/\sigma_0^2 \quad \dots(13.14)$$

follows chi-square distribution with  $(n - 1)$  d.f.

By comparing the calculated value with the tabulated value of  $\chi^2$  for  $(n - 1)$  d.f. at certain level of significance, (usually 5%), we may retain or reject the null hypothesis.

**Remarks.** 1. The above test (13.14) can be applied only if the population from which sample is drawn is normal.

2. If the sample size  $n$  is large ( $>30$ ), then we can use Fisher's approximation

$$\sqrt{2\chi^2} \sim N(\sqrt{2n-1}, 1)$$

$$\text{i.e., } Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0, 1) \quad \dots(13.14a)$$

and apply Normal Test.

3. For a detailed discussion on the *significant values*, (critical values), for testing  $H_0 : \sigma^2 = \sigma_0^2$  against various alternatives : (i)  $\sigma^2 > \sigma_0^2$ ; (ii)  $\sigma^2 < \sigma_0^2$  and (iii)  $\sigma^2 \neq \sigma_0^2$ , see Remark 1 to § 16.7.4.

**Example 13.9.** It is believed that the precision (as measured by the variance) of an instrument is no more than 0.16. Write down the null and alternative hypothesis for testing this belief. Carry out the test at 1% level, given 11 measurements of the same subject on the instrument :

2.5, 2.3, 2.4, 2.3, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, 2.5.  
[Calicut Univ. B.Sc. (Main Stat.), April 1989]

**Solution.** Null Hypothesis.  $H_0 : \sigma^2 = 0.16$

Alternative Hypothesis :  $H_1 : \sigma^2 > 0.16$ .

#### COMPUTATION OF SAMPLE VARIANCE

X	X - $\bar{X}$	$(X - \bar{X})^2$
2.5	- 0.01	0.0001
2.3	- 0.21	0.0441
2.4	- 0.11	0.0121
2.3	- 0.21	0.0441
2.5	- 0.01	0.0001
2.7	+ 0.19	0.0361
2.5	- 0.01	0.0001
2.6	+ 0.09	0.0081
2.6	+ 0.09	0.0081
2.7	+ 0.19	0.0361
2.5	- 0.01	0.0001
$\bar{X} = \frac{27.6}{11} = 2.51$		$\Sigma(X - \bar{X})^2 = 0.1891$

Under the null hypothesis  $H_0 : \sigma^2 = 0.16$ , the test statistic is :

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\sum(X - \bar{X})^2}{\sigma^2} = \frac{0.1891}{0.16} = 1.182$$

which follows  $\chi^2$ -distribution with  $df. (11 - 1) = 10$ .

Since the calculated value of  $\chi^2$  is less than the tabulated value 23.2 of  $\chi^2$  for 10 d.f. at 1% level of significance, it is not significant. Hence  $H_0$  may be accepted and we conclude that the data are consistent with the hypothesis that the precision of the instrument is 0.16.

**Example 3.10.** Test the hypothesis that  $\sigma = 10$ , given that  $s = 15$  for a random sample of size 50 from a normal population.

**Solution.** Null Hypothesis,  $H_0 : \sigma = 10$ .

We are given  $n = 50$ ,  $s = 15$

$$\therefore \chi^2 = \frac{ns^2}{\sigma^2} = \frac{50 \times 225}{100} = 112.5$$

Since  $n$  is large, using (13.14a), the test statistic is

$$Z = \sqrt{2\chi^2} - \sqrt{2n - 1} \sim N(0, 1)$$

$$\text{Now, } Z = \sqrt{225} - \sqrt{99} = 15 - 9.95 = 5.05$$

Since  $|Z| > 3$ , it is significant at all levels of significance and hence  $H_0$  is rejected and we conclude that  $\sigma \neq 10$ .

**13.7.2. Chi-square Test of Goodness of Fit.** A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as "Chi-square test of goodness of fit." It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

If  $O_i$ , ( $i = 1, 2, \dots, n$ ) is a set of observed (experimental) frequencies and  $E_i$ , ( $i = 1, 2, \dots, n$ ) is the corresponding set of expected (theoretical or hypothetical) frequencies, then Karl Pearson's chi-square, given by

$$\chi^2 = \sum_{i=1}^n \left[ \frac{(O_i - E_i)^2}{E_i} \right], \quad \left( \sum_{i=1}^n O_i = \sum_{i=1}^n E_i \right) \quad \dots (13.15)$$

follows chi-square distribution with  $(n - 1)$  d.f.

**Remark.** This is an approximate test for large values of  $n$ . The conditions for the validity of the  $\chi^2$ -test of goodness of fit have already been given in § 13.5 on page 13.15.

**Example 13.11.** The following figures show the distribution of digits in numbers chosen at random from a telephone directory :

Digits : 0 1 2 3 4 5 6 7 8 9 Total

Frequency : 1026 1107 997 966 1075 933 1107 972 964 853 10,000

Test whether the digits may be taken to occur equally frequently in the directory. [Osmania Univ. M.A. (Eco.), 1992]

**Solution.** Here we set up the *null hypothesis* that the digits occur equally frequently in the directory.

Under the null hypothesis, the expected frequency for each of the digits 0, 1, 2, ..., 9 is  $10000/10 = 1000$ . The value of  $\chi^2$  is computed as follows:

## CALCULATIONS FOR $\chi^2$

Digits	Observed Frequency (O)	Expected Frequency (E)	$(O - E)^2$	$(O - E)^2/E$
0	1026	1000	676	0.676
1	1107	1000	11449	11.449
2	997	1000	9	0.009
3	966	1000	1156	1.156
4	1075	1000	5625	5.625
5	933	1000	4489	4.489
6	1107	1000	11449	11.449
7	972	1000	784	0.784
8	964	1000	1296	1.296
9	853	1000	21609	21.609
Total	10,000	10,000		58.542

$$\therefore \chi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 58.542$$

The number of degrees of freedom =  $10 - 1 = 9$ , (since we are given 10 frequencies subjected to only one linear constraint  $\sum O = \sum E = 10,000$ ).

The tabulated  $\chi^2_{0.05}$  for 9 d.f. = 16.919

Since the calculated  $\chi^2$  is much greater than the tabulated value, it is highly significant and we reject the null hypothesis. Thus we conclude that the digits are not uniformly distributed in the directory.

**Example 13.12.** The following table gives the number of aircraft accidents that occurs during the various days of the week. Find whether the accidents are uniformly distributed over the week.

<i>Days</i>	<i>... Sun.</i>	<i>Mon.</i>	<i>Tues.</i>	<i>Wed.</i>	<i>Thus.</i>	<i>Fri.</i>	<i>Sat.</i>	
<i>No. of accidents</i>	<i>...</i>	<i>14</i>	<i>16</i>	<i>8</i>	<i>12</i>	<i>11</i>	<i>9</i>	<i>14</i>

(Given : the values of chi-square significant at 5, 6, 7, df. are respectively 11.07, 12.59, 14.07 at the 5% level of significance.

**Solution.** Here we set up the null hypothesis that the accidents are uniformly distributed over the week.

$$\begin{aligned}\chi^2 &= \frac{(14-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(12-12)^2}{12} \\ &\quad + \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(14-12)^2}{12} \\ &= \frac{1}{12}(4+16+16+0+1+9+4) = \frac{50}{12} \\ &= 4.17\end{aligned}$$

The number of degrees of freedom

$$\begin{aligned}&= \text{Number of observations} - \text{Number of independent constraints.} \\ &= 7 - 1 = 6\end{aligned}$$

The tabulated  $\chi^2_{0.05}$  for 6 d.f. = 12.59

Since the calculated  $\chi^2$  is much less than the tabulated value, it is highly insignificant and we accept the null hypothesis. Hence we conclude that the accidents are uniformly distributed over the week.

**Example 13-13.** The theory predicts the proportion of beans in the four groups A, B, C and D should be 9 : 3 : 3 : 1. In an experiment among 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory? [Agra Univ. B.Sc., 1991]

**Solution.** Null Hypothesis : We set up the null hypothesis that the theory fits well into the experiment, i.e., the experimental results support the theory.

Under the null hypothesis, the expected (theoretical) frequencies can be computed as follows :

$$\text{Total number of beans} = 882 + 313 + 287 + 118 = 1600$$

These are to be divided in the ratio 9 : 3 : 3 : 1

$$\therefore E(882) = \frac{9}{16} \times 1600 = 900, \quad E(313) = \frac{3}{16} \times 1600 = 300$$

$$E(287) = \frac{3}{16} \times 1600 = 300, \quad E(118) = \frac{1}{16} \times 1600 = 100$$

$$\begin{aligned}\therefore \chi^2 &= \sum \left[ \frac{(O-E)^2}{E} \right] \\ &= \frac{(882-900)^2}{900} + \frac{(313-300)^2}{300} + \frac{(287-300)^2}{300} + \frac{(118-100)^2}{100} \\ &= 0.3600 + 0.5633 + 0.5633 + 3.2400 = 4.7266\end{aligned}$$

$$\text{d.f.} = 4 - 1 = 3, \text{ and tabulated } \chi^2_{0.05} \text{ for 3 d.f.} = 7.815$$

Since the calculated value of  $\chi^2$  is less than the tabulated value, it is not significant. Hence the null hypothesis may be accepted at 5% level of significance and we may conclude that there is good correspondence between theory and experiment.

**Example 13-14.** A survey of 320 families with 5 children each revealed the following distribution :

No. of boys :	5	4	3	2	1	0
No. of girls :	0	1	2	3	4	5
No. of families :	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally probable?

**Solution.** Let us set up the null hypothesis that the data are consistent with the hypothesis of equal probability for male and female births. Then under the null hypothesis :

$$p = \text{Probability of male birth} = \frac{1}{2} = q$$

$$p(r) = \text{Probability of } r \text{ male births in a family of 5}$$

$$= \binom{5}{r} p^r q^{5-r} = \binom{5}{r} \left(\frac{1}{2}\right)^r$$

The frequency of  $r$  male births is given by :

$$f(r) = N \cdot p(r) = 320 \times \binom{5}{r} \times \left(\frac{1}{2}\right)^r$$

$$= 10 \times \binom{5}{r} \quad \dots (*)$$

Substituting  $r = 0, 1, 2, 3, 4$  successively in (\*), we get the expected frequencies as follows :

$$f(0) = 10 \times 1 = 10, \quad f(1) = 10 \times {}^5C_1 = 50$$

$$f(2) = 10 \times {}^5C_2 = 100, \quad f(3) = 10 \times {}^5C_3 = 100$$

$$f(4) = 10 \times {}^5C_4 = 50, \quad f(5) = 10 \times {}^5C_5 = 10$$

#### CALCULATIONS FOR $\chi^2$

Observed Frequencies (O)	Expected Frequencies (E)	$(O - E)^2$	$(O - E)^2/E$
14	10	16	1.6000
56	50	36	0.7200
110	100	100	1.0000
88	100	144	1.4400
40	50	100	2.0000
12	10	4	0.4000
Total 320	320		7.1600

$$\therefore \chi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 7.16$$

Tabulated  $\chi^2_{0.05}$  for  $6 - 1 = 5$  d.f. is 11.07.

Calculated value of  $\chi^2$  is less than the tabulated value, it is not significant at 5% level of significance and hence the null hypothesis of equal probability for male and female births may be accepted.

**Example 13-15.** Fit a Poisson distribution to the following data and test the goodness of fit.

X :	0	1	2	3	4	5	6
f:	275	72	30	7	5	2	1

**Solution.** Mean of the given distribution is :

$$\bar{X} = \frac{\sum f_i x_i}{N} = \frac{189}{392} = 0.482$$

In order to fit a Poisson distribution to the given data, we take the mean (parameter)  $m$  of the Poisson distribution equal to the mean of the given distribution, i.e., we take

$$m = \bar{X} = 0.482$$

The frequency of  $r$  successes is given by the Poisson law as :

$$f(r) = Np(r) = 392 \times \frac{e^{-0.482} (0.482)^r}{r!}; r = 0, 1, 2, \dots, 6$$

$$\begin{aligned} \text{Now } f(0) &= 392 \times e^{-0.482} = 392 \times \text{Antilog} [-0.482 \log e] \\ &= 392 \times \text{Antilog} [-0.482 \times \log 2.7183] \quad (\because e = 2.7183) \\ &= 392 \times \text{Antilog} [-0.482 \times 0.4343] \\ &= 392 \times \text{Antilog} [-0.2093] \end{aligned}$$

$$\begin{aligned} &= 392 \times \text{Antilog} [1.7907] = 392 \times 0.6176 \\ &= 242.1 \end{aligned}$$

$$f(1) = m \times f(0) = 0.482 \times 242.1 = 116.69$$

$$f(2) = \frac{m}{2} \times f(1) = 0.241 \times 116.69 = 28.12$$

$$f(3) = \frac{m}{3} \times f(2) = \frac{0.482}{3} \times 28.12 = 4.518$$

$$f(4) = \frac{m}{4} \times f(3) = \frac{0.482}{4} \times 4.518 = 0.544$$

$$f(5) = \frac{m}{5} \times f(4) = \frac{0.482}{5} \times 0.544 = 0.052$$

$$f(6) = \frac{m}{6} \times f(5) = \frac{0.482}{6} \times 0.052 = 0.004$$

Hence the theoretical Poisson frequencies correct to one decimal place are as given below :

X	0	1	2	3	4	5	6	Total
Expected Frequency	242.1	116.7	28.1	4.5	0.5	0.1	0	392

#### CALCULATIONS FOR CHI-SQUARE

Observed frequency (O)	Expected frequency (E)	(O - E)	(O - E) <sup>2</sup>	(O - E) <sup>2</sup> /E
275	242.1	32.9	1082.41	4.471
72	116.7	44.7	1998.09	17.121
30	28.1	1.9	3.61	0.128

$\left. \begin{matrix} 7 \\ 5 \\ 2 \\ 1 \end{matrix} \right\} 15$	$\left. \begin{matrix} 4.5 \\ 0.5 \\ 0.1 \\ 0 \end{matrix} \right\} 5.1$	9.9	98.01	19.217
392	392.0			40.937

$$\therefore \chi^2 = \sum \frac{(O - E)^2}{E} = 40.937$$

Degrees of freedom =  $7 - 1 - 1 - 3 = 2$

(One d.f. being lost because of the linear constraint  $\sum O = \sum E$ ; 1 d.f. is lost because the parameter  $m$  has been estimated from the given data and is then used for computing the expected frequencies; 3 d.f. are lost because of pooling the last four expected cell frequencies which are less than five.)

Tabulated value of  $\chi^2$  for 2 d.f. at 5% level of significance is 5.99.

*Conclusion.* Since calculated value of  $\chi^2$  (40.937) is much greater than 5.99, it is highly significant. Hence we conclude that Poisson distribution is not a good fit to the given data.

### EXERCISE 13(b)

1. (a) Define Chi-square and obtain its sampling distribution. Mention some prominent features of its frequency curve. Obtain the mean and the variance of the chi-square distribution.

(b) Show that the sum of two independent variates having chi-square distributions, has a chi-square distribution.

2. (a) Write a short note on the Chi-square test of goodness of fit of a random sample to a hypothetical distribution

(b) Describe the Chi-square test of significance and state the various uses to which it can be put.

(c) Discuss the  $\chi^2$ -test of goodness of fit of a theoretical distribution to an observed frequency distribution. How are the degrees of freedom ascertained when some parameters of the theoretical distribution have to be estimated from the data?

3. (a) The following table gives the number of aircraft accidents that occurred during the seven days of the week. Find whether the accidents are uniformly distributed over the week.

Days	: Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Total
No. of accidents :	14	18	12	11	15	14	84

Ans.  $H_0$  : Accidents are uniformly distributed over the week.  $\chi^2 = 2.143$ ; Not significant.  $H_0$  may be accepted.

(b) A die is thrown 60 times with the following results.

Face	: 1	2	3	4	5	6
Frequency	: 8	7	12	8	14	11

Test at 5% level of significance if the die is honest, assuming that  $P(\chi^2 > 11.1) = 0.05$  with 5 d.f. [Burdwan Univ. B.Sc. (Hons.), 1991]

4. (a) In 250 digits from the lottery numbers, the frequencies of the digits 0, 1, 2, ..., 9 were 23, 25, 20, 23, 23, 22, 29, 25, 33 and 27. Test the hypothesis that they were randomly drawn.

(b) 200 digits were chosen at random from a set of tables. The frequencies of the digits were :

Digits	:	0	1	2	3	4	5	6	7	8	9
Frequency	:	18	19	23	21	16	25	22	20	21	15

Use  $\chi^2$  test to assess the correctness of hypothesis that the digits were distributed in equal numbers in the table, given that the values of  $\chi^2$  are respectively 16.9, 18.3 and 19.7 for 9, 10 and 11 degrees of freedom at 5% level of significance.

[Delhi Univ. B.Sc., 1992]

Ans.  $\chi^2 = 4.3$ . Hypothesis seems to be correct.

5. Among 64 offsprings of a certain cross between guinea pigs, 34 were red, 10 were black and 20 were white. According to the genetic model these numbers should be in the ratio 9 : 3 : 4. Are the data consistent with the model at 5 per cent level ?

[You are given that the value of  $\chi^2$  with the probability 0.05 being exceeded is 5.99 for 2 d.f. and 3.84 for 1 d.f.]

6. In an experiment on pea-breeding, Mendal obtained the following frequencies of seeds : 315 round and yellow, 101 wrinkled and yellow; 108 round and green, 32 wrinkled and green. Total 556. Theory predicts that the frequencies should be in the proportion 9 : 3 : 3 : 1 respectively. Set up proper hypothesis and test it at 10% level of significance.

Ans.  $\chi^2 = 0.51$ . There seems to be good correspondence between theory and experiment.

7. (a) Selfed progenies of a cross between pure strains of plant segregated as follows :

	<i>Early flowering</i>	<i>Late flowering</i>
Tall	:	120
Short	:	36

Do the results agree with the theoretical frequencies which specify a 9 : 3 : 3 : 1 ratio ?

(b) Children having one parent of blood-type *M* and the other type *N* will always be one of the three types *M*, *MN*, *N* and average proportions of these will be 1 : 2 : 1.

Out of 300 children having one *M* parent and one *N* parent, 30% were found to be of type *M*, 45% of type *MN* and the remaining of type *N*. Use  $\chi^2$  to test the hypothesis.

[Patna Univ. B.Sc., 1991]

(c) A genetical law says that children having one parent of blood group *M* and the other parent of blood group *N* will always be one of the three blood groups *M*, *MN*, *N*; and that the average number of children in these groups will be in the ratio 1 : 2 : 1. The report on an experiment states as follows : "Of 162 children having one *M* parent, and one *N* parent, 28.4% were found to be of group *M*, 42% of group *MN* and the rest of the group *N*". Do the data in the report conform to the expected genetic ratio 1 : 2 : 1 ?

(d) A bird watcher sitting in a park has spotted a number of birds belonging to 6 categories. The exact classification is given below :

Category	1	2	3	4	5	6
Frequency	6	7	13	17	6	5

Test at 5% level of significance whether or not the data is compatible with the assumption that this particular park is visited by birds belonging to these six categories in the proportion  $1 : 1 : 2 : 3 : 1 : 1$ .

[Given  $P(\chi^2 \geq 11.07) = 0.05$  for 5 degrees of freedom]

[Calcutta Univ. B.Sc. (Maths. Hons.), 1991]

(e) Every clinical thermometer is classified into one of four categories  $A, B, C, D$  on the basis of inspection and test. From past experience it is known that thermometers produced by a certain manufacturer are distributed among the four categories in the following proportions :

Category	A	B	C	D
Proportion	0.87	0.09	0.03	0.01

A new lot of 1336 thermometers is submitted by the manufacturer for inspection and test and the following distribution into four categories results :

Category	A	B	C	D
No. of thermometers reported	1188	91	47	10

Does this new lot of thermometers differ from the previous experience with regards to proportion of thermometers in each category ?

8. (a) Five unbiased dice were thrown 96 times and the number of times 4, 5 or 6 was obtained is given below.

No. of dice showing 4, 5 or 6 :	5	4	3	2	1	0
Frequency	8	18	35	24	10	1

Fit a suitable distribution and test for the goodness of fit as far as you can proceed without the use of any tables and state how you would proceed further.

[Gauhati Univ. B.Sc., 1992]

(b) In 120 throws of a single die, the following distribution of faces was obtained :

Faces	1	2	3	4	5	6	Total
Frequency	30	25	18	10	22	15	120

Compute the statistic you would use to test whether the results constituted refutation of the "equal probability" (null hypothesis). Also state how you would proceed further.

[Nagpur Univ. B.Sc., 1992]

(c) Given below is the number of male and female births in 1,000 families having five children :

Number of male births	Number of female births	Number of families
0	5	40
1	4	300
2	3	250
3	2	200
4	1	30

Test whether the given data is consistent with the hypothesis that the binomial law holds if the chance of a male birth is equal to that of a female birth.

(d) Five-pig litters		Six-pig litters	
Number of males in litter	Number of litters	Number of males in litter	Number of litters
0	2	0	3
1	20	1	16
2	41	2	53
3	35	3	78
4	14	4	53
5	4	5	18
		6	0

Test whether each of the above two samples is a binomial sample (i) with  $p = 0.5$ , given *a priori* and (ii) with  $p$  determined from the data. Test the significance of the difference between the two sample  $p$ 's.

9. (a) The following table gives the count of yeast cells in square of a cyclometer. A square millimeter is divided into 400 equal squares and the number of these squares containing 0, 1, 2, ... cells are recorded—

Number of cells : 0 1 2 3 4 5 6 7 8 9 10

Frequency : 0 20 43 53 86 70 54 37 18 10 5

Number of cells : 11 12 13 14 15 16

Frequency : 2 2 0 0 0 0

Fit a Poisson distribution to the data and test the goodness of fit.

(b) The following is the distribution of the hourly number of trucks arriving at a company's warehouse.

Trucks arriving per hour : 0 1 2 3 4 5 6 7 8

Frequency : 52 151 130 102 45 12 5 1 2

Find the mean of the distribution and using its mean, (rounded to one decimal) as the parameter  $\lambda$ , fit a Poisson distribution. Test for goodness of fit at the level of significance  $\alpha = 0.05$

[Madras Institute of Technology, 1992]

(c) Obtain the equation of the normal curve that may be fitted to the following data :

Class : 60—65 65—70 70—75 75—80 80—85 85—90 90—95 95—100

Frequency : 3 21 150 335 326 135 26 4

Obtain the expected normal frequencies and test the goodness of fit.

10. Aitken gives the following distribution of times shown by two samples of 504 watches each, displayed in watch-maker's windows :

Class interval for time shown	Frequency of watches from sample I	Frequency of watches from sample II
0—2	75	83
2—4	93	86
4—6	94	94
6—8	76	72
8—10	80	82
10—12	86	87
Total	504	504

Calculate the expected frequencies of watches in the various class intervals under the hypothesis that the times shown are uniformly distributed over the interval (0, 12), separately for the two samples and also for the combined sample of all the 1,008 watches.

Test the goodness of fit for the two samples separately and for the combined sample. Test also the significance of the sum of the values of  $\chi^2$  for the two separate samples.

11. The following independent observations were made on the price of grain in 10 consecutive months :

Month	: 1	2	3	4	5	6	7	8	9	10
Price (in Rs.)	115	118	120	140	135	137	139	142	144	150

Test the hypothesis that the expected price in the  $i$ th month is Rs.  $(100 + 3i)$ ,  $i = 1, 2, \dots, 10$  with a standard deviation of Rs. 5 under the assumption that the prices are normally distributed.

12. To test a hypothesis  $H_0$ , an experiment is performed 3 times. The resulting values of chi-square are 2.37, 1.86 and 3.54, each of which corresponds to one degree of freedom. Show that while  $H_0$  cannot be rejected at 5% level on the basis of any individual experiment, it can be rejected when the three experiments are collectively counted. [Poona Univ. B.Sc., 1991]

[Hint. Use additive property of chi-square variates.]

13. (a) Describe the chi-square test for testing a hypothesis that a normal population has a specified variance  $\sigma^2$ .

(b) Give the approximation to the test statistic in (a) if  $n$ , the sample size, is sufficiently large.

(c) A sample of 15 values shows that the s.d. is 6.4. Is this compatible with the hypothesis that the sample is from a normal population with s.d. 5 ?

Ans.  $H_0 : \sigma = 5$ ,  $\chi^2 = 24.58$ ; Significant. Population s.d. is not 5.

(d) Test the hypothesis that  $\sigma = 8$ , given that  $s = 10$  for a random sample of size 51 from a normal population.

Ans.  $Z = \sqrt{2\chi^2} - \sqrt{(2n-1)} = \sqrt{2 \times 79.69} - \sqrt{101} = 2.57$ . Significant at 5% level of significance.

14. (a) A manufacturer claims that the life time of a certain brand of batteries produced by his factory has a variance of 5000 (hours)<sup>2</sup>. A sample of size 26 has a variance of 7200 (hours)<sup>2</sup>. Assuming that it is reasonable to treat these data as a random sample from a normal population, test the manufacturer's claim at the  $\alpha = 0.02$  level.

Hint.  $H_0 : \sigma^2 = 5000$  (hours)<sup>2</sup>;  $H_1 : \sigma^2 \neq 5000$  (hours)<sup>2</sup> (Two-tailed)

Critical region is :  $\chi^2 < \chi^2_{(25)} (0.99)$  ar. i  $\chi^2 > \chi^2_{(25)} (0.01)$ .

(b) A manufacturer recorded the cut-off bias (Volt) of a sample of 10 tubes as follows :

12.1, 12.3, 11.8, 12.0, 12.4, 12.0, 12.1, 11.9, 12.2, 12.2

The variability of cut-off bias for tubes of a standard type as measured by the standard deviation is 0.208 volts. Is the variability of the new tube with respect to cut-off bias less than that of the standard type ?

**Hint :**  $H_0 : \sigma^2 = (0.208)^2$  (Volts) $^2 = \sigma_0^2$  (say);  $H_1 : \sigma^2 < \sigma_0^2$

Critical region is :  $\chi^2 < \chi^2_{(n-1)} (1 - \alpha) = \chi^2_{(9)} (0.95); \alpha = 0.05$

**13-7-3. Independence of Attributes.** Let us consider two attributes  $A$  and  $B$ ,  $A$  divided into  $r$  classes  $A_1, A_2, \dots, A_r$ , and  $B$  divided into  $s$  classes  $B_1, B_2, \dots, B_s$ . Such a classification in which attributes are divided into more than two classes is known as *manifold classification*. The various cell frequencies can be expressed in the following table known as  $r \times s$  *contingency table* where  $(A_i)$  is the number of persons possessing the attribute  $A_i$ , ( $i = 1, 2, \dots, r$ ),  $(B_j)$  is the number of persons possessing the attribute  $B_j$ , ( $j = 1, 2, \dots, s$ ) and  $(A_i B_j)$  is the number of persons possessing both the attributes  $A_i$  and  $B_j$ , [ $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, s$ ]. Also

$$\sum_{i=1}^r (A_i) = \sum_{j=1}^s (B_j) = N, \text{ is the total frequency.}$$

$r \times s$  CONTINGENCY TABLE

$A$	$A_1$	$A_2$	.....	$A_i$	.....	$A_r$	Total
$B$	$(A_1 B_1)$	$(A_2 B_1)$	.....	$(A_i B_1)$	.....	$(A_r B_1)$	$(B_1)$
$B_2$	$(A_1 B_2)$	$(A_2 B_2)$	.....	$(A_i B_2)$	.....	$(A_r B_2)$	$(B_2)$
⋮							
$B_j$	$(A_1 B_j)$	$(A_2 B_j)$	.....	$(A_i B_j)$	.....	$(A_r B_j)$	$(B_j)$
⋮							
$B_s$	$(A_1 B_s)$	$(A_2 B_s)$	.....	$(A_i B_s)$	.....	$(A_r B_s)$	$(B_s)$
Total	$(A_1)$	$(A_2)$	.....	$(A_i)$	.....	$(A_r)$	$N$

The problem is to test if two attributes  $A$  and  $B$  under consideration are independent or not.

Under the *null hypothesis that the attributes are independent*, the theoretical cell frequencies are calculated as follows :

$P[A_i] = \text{Probability that a person possesses the attribute } A_i$

$$= \frac{(A_i)}{N} ; i = 1, 2, \dots, r$$

$P[B_j]$  = Probability that a person possesses the attribute  $B_j$

$$= \frac{(B_j)}{N} ; j = 1, 2, \dots, s$$

$P[A_i B_j]$  = Probability that a person possesses the attributes  $A_i$  and  $B_j$   
 $= P(A_i)P(B_j)$

(By compound probability theorem, since the attributes  $A_i$  and  $B_j$  are independent, under the null hypothesis).

$$\therefore P[A_i B_j] = \frac{(A_i)}{N} \cdot \frac{(B_j)}{N} ; i = 1, 2, \dots, r ; j = 1, 2, \dots, s$$

$\therefore (A_i B_j)_0$  = Expected number of persons possessing both the attributes  $A_i$  and  $B_j$

$$= N.P[A_i B_j] = \frac{(A_i)(B_j)}{N}$$

$$\Rightarrow (A_i B_j)_0 = \frac{(A_i)(B_j)}{N}, (i = 1, 2, \dots, r ; j = 1, 2, \dots, s) \quad \text{--- (13.16)}$$

By using this formula we can find out expected frequencies for each of the cell-frequencies  $(A_i B_j)$ , ( $i = 1, 2, \dots, r ; j = 1, 2, \dots, s$ ), under the null hypothesis of independence of attributes.

The exact test for the independence of attributes is very complicated but a fair degree of approximation is given, for large samples, (large  $N$ ), by the  $\chi^2$ -test of goodness of fit, viz.,

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \left[ \frac{((A_i B_j) - (A_i B_j)_0)^2}{(A_i B_j)_0} \right], \quad \text{--- (13.16a)}$$

which is distributed as a  $\chi^2$ -variate with  $(r-1)(s-1)$  d.f. [c.f. Note below on degrees of freedom].

**Remark.**  $\phi^2 = \chi^2/N$  is known as *mean-square contingency*.

Since the limits for  $\chi^2$  and  $\phi^2$  vary in different cases, they cannot be used for establishing the closeness of the relationship between qualitative characters under study. Prof. Karl Pearson suggested another measure, known as "coefficient of mean square contingency" which is denoted by  $C$  and is given by

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{\phi^2}{1 + \phi^2}} \quad \text{--- (13.17)}$$

Obviously  $C$  is always less than unity. The maximum value of  $C$  depends on  $r$  and  $s$ , the number of classes into which  $A$  and  $B$  are divided. In a  $r \times r$  contingency table, the maximum value of  $C = \sqrt{(r-1)/r}$ . Since the maximum value of  $C$  differs for different classification, viz.,  $r \times r$  ( $r = 2, 3, 4, \dots$ ), strictly speaking, the values of  $C$  obtained from different types of classifications are not comparable.

**Note on Degrees of Freedom (d.f.).** The number of independent variates which make up the statistic (e.g.,  $\chi^2$ ) is known as the *degrees of freedom (d.f.)* and is usually denoted by  $v$  (the letter 'Nu' of the Greek alphabet).

The number of degrees of freedom, in general, is the total number of observations less the number of independent constraints imposed on the observations. For example, if  $k$  is the number of independent constraints in a set of data of  $n$  observations then  $v = (n - k)$ .

Thus in a set of  $n$  observations usually, the degrees of freedom for  $\chi^2$  are  $(n - 1)$ , one d.f. being lost because of the linear constraint  $\sum_i O_i = \sum_i E_i = N$ , on the frequencies (c.f. Theorem 13-3, page 13-12.)

If ' $r$ ' independent linear constraints are imposed on the cell frequencies, then the d.f. are reduced by ' $r$ '.

In addition, if any of the population parameter(s) is (are) calculated from the given data and used for computing the expected frequencies then in applying  $\chi^2$ -test of goodness of fit, we have to subtract one d.f. for each parameter calculated. Thus if ' $s$ ' is the number of population parameters estimated from the sample observations ( $n$  in number), then the required number of degrees of freedom for  $\chi^2$ -test is  $(n - s - 1)$ .

If any one or more of the theoretical frequencies is less than 5 then in applying  $\chi^2$ -test we have also to subtract the degrees of freedom lost in pooling these frequencies with the preceding or succeeding frequency (or frequencies).

In a  $r \times s$  contingency table, in calculating the expected frequencies, the row totals, the column totals and the grand totals remain fixed. The fixation of ' $r$ ' column totals and ' $s$ ' row totals imposes  $(r + s - 1)$  constraints on the cell frequencies. But since

$$\sum_{i=1}^r (A_i) = \sum_{j=1}^s (B_j) = N,$$

the total number of independent constraints is only  $(r + s - 1)$ . Further, since the total number of the cell-frequencies is  $r \times s$ , the required number of degrees of freedom is :

$$v = rs - (r + s - 1) = (r - 1)(s - 1)$$

**Example 13-6.** Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the table. Examine whether the nature of the area is related to voting preference in this election.

Votes for Area	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

**Solution.** Under the *null hypothesis* that the nature of the area is independent of the voting preference in the election, we get the observed frequencies as follows :

$$E(620) = \frac{1170 \times 1000}{2000} = 585, \quad E(380) = \frac{830 \times 1000}{2000} = 415,$$

$$E(550) = \frac{1170 \times 1000}{2000} = 585, \text{ and } E(450) = \frac{830 \times 1000}{2000} = 415$$

**Aliter.** In a  $2 \times 2$  contingency table, since

$$\text{d.f.} = (2 - 1)(2 - 1) = 1,$$

only one of the cell frequencies can be filled up independently and the remaining will follow immediately, since the observed and theoretical marginal totals are fixed. Thus having obtained any one of the theoretical frequencies, (say),  $E(620) = 585$ , the remaining theoretical frequencies can be easily obtained as follows :

$$E(380) = 1000 - 585 = 415, \quad E(550) = 1170 - 585 = 585.$$

and

$$E(450) = 1000 - 415 = 585$$

$$\begin{aligned} \therefore \chi^2 &= \sum \left[ \frac{(O - E)^2}{E} \right] = \frac{(620 - 585)^2}{585} + \frac{(380 - 415)^2}{415} \\ &\quad + \frac{(550 - 585)^2}{585} + \frac{(450 - 415)^2}{415} \\ &= (35)^2 \left[ \frac{1}{585} + \frac{1}{415} + \frac{1}{585} + \frac{1}{415} \right] \\ &= (1225)[2 \times 0.002409 + 2 \times 0.001709] \doteq 10.0891 \end{aligned}$$

Tabulated  $\chi^2_{0.05}$  for  $(2 - 1)(2 - 1) = 1$  d.f. is 3.841. Since calculated  $\chi^2$  is much greater than the tabulated value, it is highly significant and null hypothesis is rejected at 5% level of significance. Thus we conclude that nature of area is related to voting preference in the election.

**Example 13-17.** ( $2 \times 2$  contingency table). For the  $2 \times 2$  table,

$a$	$b$
$c$	$d$

prove that chi-square test of independence gives

$$\chi^2 = \frac{N(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}, \quad N = a + b + c + d \quad \dots (13-18)$$

[Gauhati Univ. B.Sc., 1992]

**Solution.** Under the hypothesis of independence of attributes,

$$E(a) = \frac{(a+b)(a+c)}{N}$$

$$E(b) = \frac{(a+b)(b+d)}{N}$$

$$E(c) = \frac{(a+c)(c+d)}{N}$$

$$\text{and } E(d) = \frac{(b+d)(c+d)}{N}$$

$a$	$b$	$a+b$
$c$	$d$	$c+d$
$a+c$	$b+d$	$N$

$$\therefore \chi^2 = \frac{[a - E(a)]^2}{E(a)} + \frac{[b - E(b)]^2}{E(b)} + \frac{[c - E(c)]^2}{E(c)} + \frac{[d - E(d)]^2}{E(d)} \quad \therefore (*)$$

$$a - E(a) = a - \frac{(a+b)(a+c)}{N}$$

$$= \frac{a(a+b+c+d) - (a^2 + ac + ab + bc)}{N} = \frac{ad - bc}{N}$$

Similarly, we will get

$$b - E(b) = -\frac{ad - bc}{N} = c - E(c); d - E(d) = \frac{ad - bc}{N}$$

Substituting in (\*), we get

$$\begin{aligned} \chi^2 &= \frac{(ad - bc)^2}{N^2} \left[ \frac{1}{E(a)} + \frac{1}{E(b)} + \frac{1}{E(c)} + \frac{1}{E(d)} \right] \\ &= \frac{(ad - bc)^2}{N} \left[ \left\{ \frac{1}{(a+b)(a+c)} + \frac{1}{(a+b)(b+d)} \right\} \right. \\ &\quad \left. + \left\{ \frac{1}{(a+c)(c+d)} + \frac{1}{(b+d)(c+d)} \right\} \right] \\ &= \frac{(ad - bc)^2}{N} \left[ \frac{b+d+a+c}{(a+b)(a+c)(b+d)} + \frac{b+d+a+c}{(a+c)(c+d)(b+d)} \right] \\ &= (ad - bc)^2 \left[ \frac{c+d+a+b}{(a+b)(a+c)(b+d)(c+d)} \right]. \\ &= \frac{N(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \end{aligned}$$

**Example 13-18.** A random sample of students of Bombay University was selected and asked their opinions about 'autonomous colleges'. The results are given below. The same number of each sex was included within each class-group. Test the hypothesis at 5% level that opinions are independent of the class groupings :—

Class	Numbers		Total
	Favouring 'autonomous colleges'	Opposed to 'autonomous colleges'	
F.Y. B.A./B.Sc./B.Com.	120	80	200
S.Y. B.A./B.Sc./B.Com.	130	70	200
T.Y. B.A./B.Sc./B.Com.	70	30	100
M.A./M.Sc./M.Com.	80	20	100
Total	400	200	600

[Bombay Univ. B.Sc., April 1989]

**Solution.** We set up the null hypothesis that the opinions about autonomous colleges are independent of the class-groupings.

Here the frequencies are arranged in the form of a  $4 \times 2$  contingency table. Hence the d.f. are  $(4 - 1) \times (2 - 1) = 3 \times 1 = 3$ . Hence we need to compute independently only three expected frequencies and the remaining expected frequencies can be obtained by subtraction from the row and column totals.

Under the null hypothesis of independence :

$$E(120) = \frac{400 \times 200}{600} = 133.33 ; E(130) = \frac{400 \times 200}{600} = 133.33$$

$$E(70) = \frac{400 \times 100}{600} = 66.67$$

Now the table of expected frequencies can be completed as shown below :

Class	Number		Total
	Favouring 'autonomous colleges'	Opposed to 'autonomous colleges'	
F.Y.B.A./B.Sc./B.Com.	133.33	$200 - 133.33 = 66.67$	200
S.Y.B.A./B.Sc./B.Com.	133.33	$200 - 133.33 = 66.67$	200
T.Y.B.A./B.Sc./B.Com.	66.67	$100 - 66.67 = 33.33$	100
M.A./M.Sc./M.Com.	66.67	$100 - 66.67 = 33.33$	100
Total	400	-200	600

## CALCULATIONS FOR CHI-SQUARE

<i>O</i>	<i>E</i>	<i>O - E</i>	$(O - E)^2$	$(O - E)^2/E$
120	133.33	-13.33	177.6889	1.3327
130	133.33	-3.33	11.0889	0.0832
70	66.67	3.33	11.0889	0.1663
80	66.67	13.33	177.6889	2.6652
80	66.67	13.33	177.6889	2.6652
70	66.67	3.33	11.0889	0.1663
30	33.33	-3.33	11.0889	0.3327
20	33.33	-13.33	177.6889	5.3312
Total 400	400			12.7428

$$\therefore \chi^2 = \sum \frac{(O - E)^2}{E} = 12.7428$$

Tabulated (critical) value of  $\chi^2$  for  $(4 - 1) \times (2 - 1) = 3$  d.f. at 5% level of significance is 7.815.

**Conclusion.** Since calculated value of  $\chi^2$  is greater than the tabulated value, it is significant at 5% level of significance and we reject the null hypothesis. Hence, we conclude that the opinions about autonomous colleges' are dependent on the class-groupings.

**Example 13-19.** Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows :

Researcher	No. of students in each level				Total
	Below Average	Average	Above Average	Genius	
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different? (Given 5% value of  $\chi^2$  for 2 d.f. and 3 d.f. are 5.991 and 7.82 respectively.)

**Solution.** We set up the null hypothesis that the data obtained are independent of the sampling techniques adopted by the two researchers. In other words, the null hypothesis is that there is no significant difference between the sampling techniques used by the two researchers for collecting the required data.

Here we have a  $4 \times 2$  contingency table and  $d.f. = (4 - 1) \times (2 - 1) = 3 \times 1 = 3$ . Hence we need to compute only 3 independent expected frequencies and the remaining expected frequencies can be obtained by subtraction from the marginal row and column totals.

Under the null hypothesis of independence, we have

$$E(86) = \frac{126 \times 200}{300} = 84; E(60) = \frac{93 \times 200}{300} = 62;$$

$$E(44) = \frac{69 \times 200}{300} = 46$$

The table of expected frequencies can now be completed as shown below :

Researcher	No. of students in each level				Total
	Below Average	Average	Above average	Genius	
X	84	62	46	$200 - 192 = 8$	200
Y	$126 - 84 = 42$	$93 - 62 = 31$	$69 - 46 = 23$	$12 - 8 = 4$	100
Total	126	93	69	12	300

Since we cannot apply the  $\chi^2$ -test straightway here as the last frequency is less than 5, we should use the technique of pooling in this case as given below :

#### CALCULATIONS FOR CHI-SQUARE

O	E	O - E	$(O - E)^2$	$(O - E)^2/E$
86	84	2	4	0.048
60	62	-2	4	0.064
44	46	-2	4	0.087
10	8	2	4	0.500
40	42	-2	4	0.095
33	31	2	4	0.129
25	23	0	0	0
2	4			
Total	300	0	0	0.923

$$\text{After pooling, } \chi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 0.923$$

and the  $d.f. = (4 - 1) \times (2 - 1) - 1 = 3 - 1 = 2$ , since 1 d.f. is lost in the method of pooling.

Tabulated value of  $\chi^2$  for 2d.f. at 5% level of significance is 5.991.

**Conclusion.** Since calculated value is less than the tabulated value, null hypothesis may be accepted at 5% level of significance and we may conclude that there is no significant difference in the sampling techniques used by the two researchers.

**13.8. Yates' Correction.** In a  $2 \times 2$  contingency table, the number of *df.* is  $(2 - 1)(2 - 1) = 1$ . If any one of the theoretical cell frequencies is less than 5, then the use of pooling method for  $\chi^2$ -test results in  $\chi^2$  with 0 *df.* (since 1 *df.* is lost in pooling) which is meaningless. In this case we apply a correction due to F. Yates (1934), which is usually known as "Yates' Correction for Continuity". [As already pointed out,  $\chi^2$  is a continuous distribution and it fails to maintain its character of continuity if any of the expected frequency is less than 5; hence the name 'Correction for Continuity']. This consists in adding 0.5 to the cell frequency which is less than 5 and then adjusting for the remaining cell frequencies accordingly. The  $\chi^2$ -test of goodness of fit is then applied without pooling method.

For a  $2 \times 2$  contingency table,

<i>a</i>	<i>b</i>
<i>c</i>	<i>d</i>

, we have

$$\chi^2 = \frac{N(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

According to Yate's correction, as explained above, we subtract (or add)  $\frac{1}{2}$  from *a* and *d* and add (subtract)  $\frac{1}{2}$  to *b* and *c* so that the marginal totals are not disturbed at all. Thus, corrected value of  $\chi^2$  is given as

$$\chi^2 = \frac{N[(a + \frac{1}{2})(d + \frac{1}{2}) - (b + \frac{1}{2})(c + \frac{1}{2})]^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$\text{Numerator} = N[(ad - bc) \mp \frac{1}{2}(a + b + c + d)]^2$$

$$= N \left[ |ad - bc| - \frac{N}{2} \right]^2$$

$$\therefore \chi^2 = \frac{N[|ad - bc| - N/2]^2}{(a+c)(b+d)(a+b)(c+d)} \quad \dots (13.18a)$$

**Remarks 1.** If *N* is large, the use of Yate's correction will make very little difference in the value of  $\chi^2$ . If, however, *N* is small, the application of Yates' correction may overstate the probability.

2. It is recommended by many authors and it seems quite logical in the light of the above discussion that Yates' correction be applied to every  $2 \times 2$  table, even if no theoretical cell frequency is less than 5.

**13.9. Brandt and Snedecor Formula for  $2 \times k$  Contingency Table.** Let the observations  $a_{ij}$ , ( $i = 1, 2$ ;  $j = 1, 2, \dots, k$ ) be arranged in a  $2 \times k$  contingency table as follows :

$B$	$A$	$A_1$	$A_2$	.....	$A_i$	.....	$A_k$	Total
$B_1$		$a_{11}$	$a_{12}$	.....	$a_{1i}$	.....	$a_{1k}$	$m_1$
$B_2$		$a_{21}$	$a_{22}$	.....	$a_{2i}$	.....	$a_{2k}$	$m_2$
Total		$n_1$	$n_2$	.....	$n_i$	.....	$n_k$	$N$

Under the hypothesis of independence of attributes, we have

$$E(a_{1i}) = \frac{n_i \times m_1}{N}; E(a_{2i}) = \frac{n_i \times m_2}{N}, i = 1, 2, \dots, k$$

$$\begin{aligned} \therefore \chi^2 &= \sum_{i=1}^k \left[ \frac{(a_{1i} - E(a_{1i}))^2}{E(a_{1i})} + \frac{(a_{2i} - E(a_{2i}))^2}{E(a_{2i})} \right] \\ &= \sum_{i=1}^k \left\{ \frac{\left( a_{1i} - \frac{m_1 n_i}{N} \right)^2}{\left( \frac{m_1 n_i}{N} \right)} + \frac{\left( a_{2i} - \frac{m_2 n_i}{N} \right)^2}{\left( \frac{m_2 n_i}{N} \right)} \right\} \\ &= \sum_{i=1}^k \left[ \frac{N n_i}{m_1} \left( \frac{a_{1i}}{n_i} - \frac{m_1}{N} \right)^2 + \frac{N n_i}{m_2} \left( \frac{a_{2i}}{n_i} - \frac{m_2}{N} \right)^2 \right] \\ &= \sum_{i=1}^k \left[ \frac{n_i}{p} \left( p_i - p \right)^2 + \frac{n_i}{q} \left( q_i - q \right)^2 \right] \end{aligned}$$

where  $p_i = \frac{a_{1i}}{n_i}, q_i = 1 - p_i = \frac{a_{2i}}{n_i}$

and  $p = \frac{m_1}{N}, q = 1 - p = \frac{m_2}{N}$

$$\begin{aligned} \therefore \chi^2 &= \sum_{i=1}^k \left[ \frac{n_i}{p} \left( p_i - p \right)^2 + \frac{n_i}{q} \left\{ \left( 1 - p_i \right) - \left( 1 - p \right) \right\}^2 \right] \\ &= \sum_{i=1}^k n_i \left( p_i - p \right)^2 \left\{ \frac{1}{p} + \frac{1}{q} \right\} \\ &= \sum_{i=1}^k n_i \left( p_i - p \right)^2 \frac{1}{pq} \quad [\because p + q = 1] \\ &= \frac{1}{pq} \sum_{i=1}^k n_i \left( p_i^2 + p^2 - 2p_i p \right) \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{pq} \left[ \sum_{i=1}^k n_i p_i^2 + p^2 \sum_{i=1}^k n_i - 2p \sum_{i=1}^k p_i n_i \right] \\
 &= \frac{1}{pq} \left[ \sum_{i=1}^k n_i p_i^2 + p^2 N - 2p \cdot N p \right] \\
 &= \frac{1}{pq} \left[ \sum_{i=1}^k n_i p_i^2 - N p^2 \right]
 \end{aligned}$$

But  $\sum_{i=1}^k n_i p_i^2 = \sum_{i=1}^k n_i p_i \cdot p_i = \sum_{i=1}^k a_{1i} p_i$

$$\begin{aligned}
 \therefore \chi^2 &= \frac{1}{pq} \left[ \sum_{i=1}^k a_{1i} p_i - N p^2 \right] = \frac{1}{pq} \left[ \sum_{i=1}^k \frac{a_{1i}^2}{n_i} - N p^2 \right] \quad \dots(13-19) \\
 &= \frac{1}{pq} \left[ \sum_{i=1}^k \frac{a_{1i}^2}{n_i} - m_1 p \right] = \frac{1}{pq} \left[ \sum_{i=1}^k \frac{a_{1i}^2}{n_i} - \frac{m_1^2}{N} \right] \quad \dots[13-19(a)]
 \end{aligned}$$

**Example 13-20.** The following table shows three age groups of boys and girls, (a) the number of children affected by a non-infectious disease and (b) the total number of children exposed to risk.

	Boys			Girls		
	I	II	III	I	II	III
(a)	60	25	48	96	18	42
(b)	240	470	350	530	200	210

(i) Test whether there are differences between the incidence rates in the three age groups of boys.

(ii) Test whether the boys and girls are equally susceptible or not.

**Solution.** (i) We set up the null hypothesis ( $H_0$ ) that there is no significant difference between the incidence rates in the three age-groups of boys.

In the notations of § 13-9, we have

$$\begin{aligned}
 a_{11} &= 60, a_{12} = 25, a_{13} = 48, m_1 = 133 \\
 n_1 &= 240, n_2 = 470, n_3 = 350, N = 1060
 \end{aligned}$$

$$p = \frac{m_1}{N} = \frac{133}{1060} = 0.1255, q = 1 - p = 0.8745$$

Substituting these values in (13-19a), we get

$$\begin{aligned}
 \chi^2 &= \frac{1}{(0.1255)(0.8745)} \left[ 15.00 + 1.33 + 6.58 - 133 \times 0.1255 \right] \\
 &= \frac{6.2187}{0.1097} = 56.688
 \end{aligned}$$

Here  $v = (3 - 1)(2 - 1) = 2$ .

The tabulated value of  $\chi^2$  for 2 degrees of freedom at 5% level of significance is 5.991. Since calculated value of  $\chi^2$  is much greater than tabulated value, we reject the null hypothesis and conclude that the incidence rates in the three age-groups of boys differ significantly.

(ii) Here we set up the *null hypothesis that the boys and girls are equally susceptible to the disease*. In the usual notations, we have

$$a_{11} = 60 + 25 + 48 = 133 \text{ and } a_{12} = 96 + 18 + 42 = 156, m_1 = 289$$

$$n_1 = 240 + 470 + 350 = 1060 \text{ and } n_2 = 530 + 200 + 210 = 940, N = 2000$$

$$\therefore p = \frac{289}{2000} = 0.1445, q = 1 - p = 0.8555$$

$$Np^2 = 289 \times 0.1445 = 41.76$$

$$\therefore \chi^2 = \frac{1}{(0.1445)(0.8555)} [16.69 + 25.89 - 41.76] = 6.605$$

$$\text{Here } v = (2 - 1)(2 - 1) = 1$$

From the tables, the value of  $\chi^2$  for 1 degree of freedom at 5% level of significance is 3.841 which is much less than the calculated value. We, therefore, reject the null hypothesis and conclude that boys and girls are not equally susceptible to the disease.

**Example 13.21.** Two samples of sizes  $N_1, N_2$  have respectively frequencies  $f_1, f_2, \dots, f_n$  and  $f'_1, f'_2, \dots, f'_n$  under the same headings. Show that  $\chi^2$  for such a distribution is equal to

$$\sum_{r=1}^n N_1 N_2 \left[ \frac{\left( \frac{f_r}{N_1} - \frac{f'_r}{N_2} \right)^2}{f_r + f'_r} \right]$$

[Allahabad Univ. B.Sc., 1992]

**Solution.** The  $2 \times n$  contingency table for which  $\chi^2$  is to be calculated is given below :

A B	$A_1$	$A_2$	...	$A_r$	...	$A_n$	Total
$B_1$	$f_1$	$f_2$	...	$f_r$	...	$f_n$	$N_1$
$B_2$	$f'_1$	$f'_2$	...	$f'_r$	...	$f'_n$	$N_2$

Under the hypothesis of independence of attributes, we have

$$E(f_r) = \frac{N_1(f_r + f'_r)}{N_1 + N_2}, E(f'_r) = \frac{N_2(f_r + f'_r)}{N_1 + N_2}$$

$$\chi^2 = \sum_{r=1}^n \left[ \frac{(f_r - E(f_r))^2}{E(f_r)} + \frac{(f'_r - E(f'_r))^2}{E(f'_r)} \right]$$

$$\begin{aligned}
 &= \sum_{r=1}^n \left[ \frac{(N_1 + N_2)}{N_1(f_r + f'_r)} \cdot \left\{ f_r - \frac{N_1(f_r + f'_r)}{N_1 + N_2} \right\}^2 \right. \\
 &\quad \left. + \frac{(N_1 + N_2)}{N_2(f_r + f'_r)} \left\{ f'_r - \frac{N_2(f_r + f'_r)}{N_1 + N_2} \right\}^2 \right] \\
 &= \sum_{r=1}^n \left[ \frac{(N_2 f_r - N_1 f'_r)^2}{N_1(N_1 + N_2)(f_r + f'_r)} + \frac{(N_1 f'_r - N_2 f_r)^2}{N_2(N_1 + N_2)(f_r + f'_r)} \right] \\
 &= \sum_{r=1}^n \frac{(N_2 f_r - N_1 f'_r)^2}{(N_1 + N_2)(f_r + f'_r)} \left[ \frac{1}{N_1} + \frac{1}{N_2} \right] \\
 &= \sum_{r=1}^n \left[ \frac{N_1 N_2}{f_r + f'_r} \left( \frac{f_r}{N_1} - \frac{f'_r}{N_2} \right)^2 \right]
 \end{aligned}$$

### EXERCISE 13(c)

1. (a) What is contingency table? Describe how the  $\chi^2$  distribution may be used to test whether the two criteria of classification in an  $m \times n$  contingency table are independent.

(b) State the hypothesis you test using the Chi-square statistic in a contingency table.

(c) Describe the  $\chi^2$  test for independence of attributes, stating clearly the conditions for the validity. Give a rule for calculating the number of degrees of freedom to be assigned to  $\chi^2$ . Illustrate your answer with an  $m \times n$  contingency table explaining the null hypothesis that is being tested.

2. Of 'A' candidates taking a certain paper, 'a' are successful, of 'B' taking another paper, 'b' are successful. Show how the significance of the difference between the ratios  $a/A$ ,  $b/B$  may be tested (i) by a  $\chi^2$  test on contingency table and (ii) by comparing the difference with its standard error assessed by means of a binomial distribution. (You may assume all frequencies are sufficiently large.) Prove algebraically that the value of  $\chi^2$  is the square of the ratio of  $(a/A - b/B)$  to its standard error.

3. (a) Show that for the entries in the following  $2 \times r$  contingency table,

	$A_1$	$A_2$		$A_i$		$A_r$	Total
$B_1$	$a_1$	$a_2$	.....	$a_i$	.....	$a_r$	$a$
$B_2$	$b_1$	$b_2$	.....	$b_i$	.....	$b_r$	$b$
Total	$n_1$	$n_2$	.....	$n_i$	.....	$n_r$	$n$

the value of  $\chi^2$  is

$$\chi^2 = \sum_{i=1}^r \omega_i (p_i - \bar{p})^2$$

where  $p_i = \frac{a_i}{n_i}$ ,  $p = \frac{a}{n}$ ,  $\omega_i = \frac{n_i}{pq}$  and  $q_i = 1 - p_i$ ,  $q = \frac{b}{n}$

[Madurai Univ. B.Sc., Oct. 1988]

(b) Given  $\chi^2$  contingency table representing two independent samples :

	$\mu_1$	$\mu_2$	...	$\mu_r$	Total
Sample I	$v_1$	$v_2$	...	$v_r$	$n$
Sample II	$\mu_1 + v_1$	$\mu_2 + v_2$	...	$\mu_r + v_r$	$m + n$

show that

$$\chi^2 = \frac{1}{\omega(1-\omega)} \left[ \sum_{i=1}^r \mu_i \omega_i - m \omega \right]$$

where  $\omega_i = \frac{\mu_i}{\mu_i + v_i}$  and  $\omega = \frac{m}{m+n}$ ,

can be used to test whether the samples are drawn from the sample population. Clearly state the underlying assumptions, and give the number of degrees of freedom.

(c) In a  $2 \times 3$  contingency table if  $N = x + y + z$ ,  $N' = x' + y' + z'$  and  $N = N'$ , show that

$$\chi^2 = \frac{(x-x')^2}{x+x'} + \frac{(y-y')^2}{y+y'} + \frac{(z-z')^2}{z+z'}$$

[Poona Univ. B.Sc., 1990]

(d) Show that for a  $2 \times 2$  table, the value of  $\chi^2$ , after applying Yates' correction for continuity is

$$\frac{N}{D} \left( ad - bc - \frac{N}{2} \right)^2 \quad \text{or} \quad \frac{N}{D} \left( ad - bc + \frac{N}{2} \right)^2$$

according as  $ad - bc > 0$  or  $< 0$  respectively, where

$$D = (a+b)(a+c)(b+d)(c+d).$$

(e) What is Yates' correction ? Show that for a  $2 \times 2$  contingency table, the value of  $\chi^2$  after applying this correction is :

$$\chi^2 = \frac{N[|ad - bc| - N/2]^2}{(a+b)(a+c)(b+d)(c+d)}$$

[Marathwada Univ. M.Sc., 1991]

4. Consider the following  $2 \times 2$  table of observed frequencies based on random samples (with replacement) of sizes  $n_{.1}$  and  $n_{.2}$  from two populations :

	Population I	Population II	Total
Class A	$n_{11}$	$n_{12}$	$n_{1.}$
Class B	$n_{21}$	$n_{22}$	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n$

(i) Define the  $\chi^2$ -statistic to be used for test of homogeneity of the two populations.

(ii) Show that

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{(n_{.1}n_{.2})}$$

(iii) Let

$$u = \frac{n_{11}}{n_1} - \frac{n_{12}}{n_2}$$

Calculate the mean and variance of  $u$  and indicate how you may estimate them.

5. (a) In an epidemic of certain disease 92 children contracted the disease. Of these, 41 received no treatment and of these 10 showed after-effects. Of the remainder who did receive treatment, 17 showed after-effects. Test the hypothesis that treatment was not effective.

(b) Can vaccination be regarded as a preventive measure of small-pox as evidenced by the following data ?

"Of 1482 persons exposed to smallpox in a locality, 368 in all were attacked. Of these 1482 persons, 343 were vaccinated and of these, only 35 were attacked".

6. (a) Define  $\chi^2$ . Cite some statistical problems where you can apply  $\chi^2$  for testing statistical hypothesis.

In an experiment on immunization of cattle from tuberculosis the following results were obtained :

	Affected	Unaffected
Inoculated	12	28
Not inoculated	13	7

Examine the effect of vaccine in controlling the incidence of the disease.

(b) What are contingency tables ? What is tested there ? Explain the test procedure therein.

The following data is collected on two characters :

	Cinegoers	Non-cinegoers
Literate	83	57
Illiterate	45	68

Based on this, can you conclude that there is no relation between the habit of cinema going and literacy ?

7. (a) To find whether a certain vaccination prevents a certain disease or not, an experiment was conducted and the following figures in various classes were obtained, A showing vaccination and B attacked by the disease.

	A	$\alpha$	Total
B	69	10	79
$\beta$	91	30	121
Total	160	40	200

Using  $\chi^2$ -test, analyse the results of the experiment for independence between A and B; examine whether Yate's correction modifies the conclusion or not. Test also the significance of the difference between the proportions of persons attacked by the disease among vaccinated and non-vaccinated which are  $69/160$  and  $10/40$ .

(b) A theory in finance known as Random Walk Theory suggests that short term changes in stock prices follow a random pattern. According to this theory,

yesterday's price change can tell us virtually nothing of value about to-day's price change. Let us denote the change in price of a stock on day  $t$  by  $\Delta P_t$ , and the change on the next day by  $\Delta P_{t+1}$ . Suppose we observe price changes of 240 stocks that have been randomly selected and obtain the results shown in the table below :

	$\Delta P_t > 0$	$\Delta P_t \leq 0$	Total
$\Delta P_{t+1} > 0$	47	53	100
$\Delta P_{t+1} \leq 0$	63	77	140
Total	110	130	240

Test the hypothesis that the change in stock price on day  $(t + 1)$  is independent of that on day  $t$ .

[Delhi Univ. M.A. (Eco.), 1987]

8. (a) Show that the value of  $\chi^2$  for  $2 \times 2$  contingency table

$a$	$b$
$c$	$d$

$$\text{is } \chi^2 = \frac{N(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)},$$

where  $N = a + b + c + d$ .

(b) Let  $X$  and  $Y$  denote the number of successes and failures respectively in  $n$  independent Bernoulli trials with  $p$  as the probability of success in each trial. Show that

$$\frac{(X - np)^2}{np} + \frac{(Y - n(1-p))^2}{n(1-p)}$$

can be approximated by a chi-square distribution with one degree of freedom when  $n$  is large.

[Delhi Univ. M.A. (Eco.), 1986]

9. Show that for  $r \times s$  contingency table :

- (a) Number of degrees of freedom is  $(r-1) \times (s-1)$
- (b)  $\chi^2 = N(s-1)$  or  $\chi^2 = N(r-1)$ , whichever is less
- (c)  $E(\chi^2) = N(r-1)(s-1)/(N-1)$
- (d)  $\max(C) = [(s-1)/s]^{1/2}$ ,  $r = s$ ,

where  $C$  is the coefficient of contingency and  $N$  is the total frequency.

10. (a) 1072 college students were classified according to their intelligence and economic conditions. Test whether there is any association between intelligence and economic conditions.

	Intelligence			
	Excellent	Good	Mediocre	Dull
Economic Conditions	Good	48	199	181
	Not good	81	185	190

(b) Below is given the distribution of hair colours for either sex in a university :

<i>Hair colour</i>	(1) <i>Fair</i>	(2) <i>Red</i>	(3) <i>Medium</i>	(4) <i>Dark</i>	(5) <i>Jet black</i>	Total
<i>Boys</i>	592	119	849	504	36	2100
<i>Girls</i>	544	97	677	451	14	1783
<i>Total</i>	1136	216	1526	955	50	3883

Test the homogeneity of hair colour for either sex. If the result is significant at 5 per cent level, explain the reason why it should be so.

11. (a) The following data are for a sample of 300 car owners who were classified with respect to age and the number of accidents they had during the past two years. Test whether there is any relationship between these two variables.

Accidents				
	0	1 or 2	3 or more	
Age	$\leq 21$	8	23	14
	22 — 26	21	42	12
	$\geq 27$	71	90	19

(b) For the data in the following table, test for independence between a person's ability in Mathematics and interest in Economics.

Ability in Mathematics			
Interest in Economics	Low	Average	High
	63	42	15
	58	61	31
	14	47	29

State clearly the assumptions underlying your test procedure.

[Delhi Univ. M.A. (Eco.), 1988]

12. The following table gives for a sample of married women, the level of education and marriage adjustment score :

Marriage—adjustment score

Level of Education	College	Very low	Low	High	Very high
		24	97	62	58
Education	High school	22	28	30	41
	Middle School	32	10	11	20

Can you conclude from the above, 'the higher the level of education, the greater is the degree of adjustment in marriage' ?

13. (a) The table below shows results of a survey in which 250 respondents were categorized according to level of education and attitude towards students' demonstrations at a certain college. Test the hypothesis that the two criteria of classification are independent. Let  $\alpha = 0.05$ .

Education	Attitude		
	Against	Neutral	For
Less than high school	40	25	5
High school	40	20	5
Some college	30	15	30
College graduate	15	15	10

(b) Test the hypothesis that there is no difference in the quality of the four kinds of tyres A, B, C and D based on the data given below. Use 5% level of significance.

	Tyre Brand			
	A	B	C	D
Failed to last 40,000 kms.	26	23	15	32
Lasted from 40,000 kms. to 60,000 kms.	118	93	116	121
Lasted more than 60,000 kms.	56	84	69	47

[Bangalore Univ. B.E., 1992]

(c) The results of a survey regarding radio listeners' preference for different types of music are given in the following table, with listeners classified by age group. Is preference of type of music influenced by age?

Type of music preferred	Age group		
	19—25	26—35	Above 36
National music	80	60	9
Foreign music	210	325	44
Indifferent	16	45	132

14. (a) If  $x_1, x_2, \dots, x_k$  represent the respective number of successes in  $k$  samples each of  $n$  trials, by considering a suitable  $2 \times k$  contingency table, derive an expression for  $\chi^2$  to test the homogeneity of this data.

(b) It was decided to check the dental health of children in 8 districts of a town. The condition of the teeth of 36 children from each district was examined and classified as either good or poor. The number of children with teeth in a poor condition from each of the districts was 9, 14, 12, 18, 7, 10, 15, 11. Can it be concluded that the dental health of children does not vary between districts?

13.9.1.  $\chi^2$ -test of Homogeneity of Correlation Coefficients. Let  $r_1, r_2, \dots, r_k$  be  $k$  estimates of correlation coefficients from independent samples of sizes  $n_1, n_2, \dots, n_k$  respectively.

We want to test the hypothesis that these sample correlation coefficients are the estimates of the same correlation coefficient  $\rho$  from a bivariate normal population.

Obtain the values of  $z_1, z_2, \dots, z_k$  from the Table of Fisher's  $z$ -transformation or from

$$z_i = \frac{1}{2} \log_e \left( \frac{1 + r_i}{1 - r_i} \right) = \tanh^{-1} r_i; \quad i = 1, 2, \dots, k \quad \dots(13.20)$$

These  $z_i$ 's are normally distributed about a common mean

$$\xi = \frac{1}{2} \log_e \left( \frac{1 + \rho}{1 - \rho} \right) \text{ and variance} = \frac{1}{n_i - 3} \quad \dots(13.21)$$

The minimum variance estimate  $\bar{z}$  of the common mean  $\xi$  of  $Z$ 's is obtained by weighting the values  $z_i$ 's inversely with their respective variances. The estimate of  $z$  is, therefore,

$$\bar{z} = \frac{\sum_i z_i (n_i - 3)}{\sum_i (n_i - 3)} \quad (\text{c.f. } \S\ 14.7.2)$$

so that  $(z_i - \bar{z}) \sqrt{n_i - 3}; i = 1, 2, \dots, k$  are independent standard normal variates. Hence  $\sum_{i=1}^k (n_i - 3) (z_i - \bar{z})^2$  is a  $\chi^2$ -variate with  $(k - 1)$  d.f. [By additive property of  $\chi^2$ -distribution, one d.f. being lost since  $z$  has been determined from the data.]

If  $\chi^2$  value thus obtained is greater than 5 per cent value of  $\chi^2$  for  $(k - 1)$  d.f., the hypothesis of homogeneity of correlation coefficients is rejected. If not, the correlation coefficients are supposed to be homogeneous in which case we combine the sample correlation coefficients to find the estimate  $\hat{\rho}$  of the population correlation coefficient  $\rho$ .

$$\begin{aligned} \text{We have} \quad & \bar{z} = \frac{1}{2} \log \left( \frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right) \\ \Rightarrow \quad & (1 + \hat{\rho}) = (1 - \hat{\rho}) e^{2\bar{z}} \\ \Rightarrow \quad & (1 + e^{2\bar{z}}) \hat{\rho} = e^{2\bar{z}} - 1 \\ \Rightarrow \quad & \hat{\rho} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1} = \tanh \bar{z} \end{aligned} \quad \dots(13.22)$$

**Remark.** For testing the homogeneity of independent estimates of the parent partial correlation coefficient, the above formulae hold with the only difference that for a partial correlation coefficient of order  $s$ ,  $n_i$  will be replaced by  $n_i - s$ .

**Example 13.22.** The correlation coefficient between daily ration of green grass and rate of growing calves on the basis of observations taken on 10, 14, 16, 20, 25 and 28 cows at six farms were found to be 0.318, 0.106, 0.253, 0.340, 0.116 and 0.112. Can these be considered homogeneous? If so, estimate the common correlation coefficient.

**Solution.**  $H_0$  : The given values of sample correlation coefficients are homogeneous or the samples are from equally correlated populations.

Using (13-20), we get

$$\begin{aligned} z_1 &= 0.3294, & z_2 &= 0.1063, & z_3 &= 0.2586 \\ z_4 &= 0.3541, & z_5 &= 0.1165, & \text{and} & z_6 = 0.1125 \end{aligned}$$

$$\therefore \bar{z} = \sum_i z_i (n_i - 3) / \sum_i (n_i - 3) = 0.1919$$

$$\text{Now } \chi^2 = \sum (n_i - 3)(z_i - \bar{z})^2 = 0.1008$$

Tabulated value of  $\chi^2$  for  $(6 - 1) = 5$ , degrees of freedom at 5% level of significance is 11.070.

Since the calculated value is less than the tabulated value, we may accept the null hypothesis that the sample correlation coefficients are homogeneous.

If  $\hat{\rho}$  is the pooled estimate of the population correlation coefficient, then using (13-22), we get

$$\hat{\rho} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1} = \frac{1.468 - 1}{1.468 + 1} = 0.1894$$

### 13-10. Bartlett's Test for Homogeneity of Several Independent Estimates of the Same Population Variance. Let

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2, \quad (i = 1, 2, \dots, k)$$

be the unbiased estimate of the population variance, obtained from the  $i$ th sample  $X_{ij}$ , ( $j = 1, 2, \dots, n_i$ ) and based on  $v_i = (n_i - 1)$  degr es of freedom, all the  $k$  samples being independent.

Under the null hypothesis that the samples come from the same population with variance  $\sigma^2$ , i.e., the independent estimates  $S_i^2$ , ( $i = 1, 2, \dots, k$ ) of  $\sigma^2$  are homogeneous, Bartlett proved that the statistic

$$\chi^2 = \sum_{i=1}^k \left( v_i \log \frac{S_i^2}{S^2} \right) / \left[ 1 + \frac{1}{3(k-1)} \left\{ \sum_i \left( \frac{1}{v_i} \right) - \frac{1}{v} \right\} \right] \quad \dots(13-23)$$

$$\text{where } S^2 = \frac{\sum v_i S_i^2}{\sum v_i} = \frac{\sum v_i S_i^2}{v}, \quad \sum_{i=1}^k v_i = v \quad \dots(*)$$

follows chi-square distribution with  $(k - 1)$  degrees of freedom.

**Remarks 1.**  $S^2$  defined in (\*) is also an unbiased estimate of  $\sigma^2$ , since

$$E(S^2) = \frac{\sum v_i E(S_i^2)}{\sum v_i} = \frac{(\sum v_i) \sigma^2}{\sum v_i} = \sigma^2$$

**2.** Let  $S_i^2$  and  $S_j^2$ ;  $i \neq j$ ,  $1 \leq (i, j) \leq k$  be the smallest and the largest values of the estimates respectively. If on the basis of F-test (c.f. Chapter 14), these do not differ significantly, then all the estimates  $S_i^2$  which lie between  $S_i^2$  and  $S_j^2$  won't differ significantly either and consequently all the estimates can be

reasonably regarded as homogeneous, coming from the same population. In this case, therefore, there is no need to apply Bartlett's test.

**13-11.  $\chi^2$ -Test for Pooling the Probabilities from Independent Tests to give a Single Test of Significance ( $P_\lambda$ -Test).** See Examples 13-1 and 13-2 for detailed discussion.

### EXCERCISE 13 (d)

1. Define  $\chi^2$ -statistic. What are its uses? What is Fisher's  $z$ -transformation for correlation coefficient and what are its properties? How is it used to combine the correlation coefficients between two random variables computed independently from different sources.

2. Explain the use of the chi-square statistic for testing the homogeneity of several independent estimates of population correlation coefficient, clearly stating the underlying assumptions.

3. (a) The correlation coefficients between wing length and tongue length were estimated from 2 samples each of size 44 to be 0.731 and 0.690. Test whether the correlation coefficients are significantly different or not. If not, obtain the best estimate of the common correlation coefficient.

(b) Test for equality of the correlation co-efficients between the scores in two halves of a psychological test applied to different groups of sizes 30, 20 and 25 if the corresponding sample values are 0.63, 0.48, 0.71, respectively.

4. (a) Independent samples of 21, 30, 39, 26 and 35 pairs of values yielded correlation coefficients 0.39, 0.61, 0.43, 0.54 and 0.48, respectively. Can these estimates be regarded as homogeneous? If so, find an estimate of the correlation coefficient in the population.

(b) Test whether the following set of correlation coefficients between stature and sitting heights obtained for persons from 8 districts can be regarded as homogeneous.

Sample size      130    60    338    78    125    299    170    139

Corr. coefficient : 0.718 0.961 0.825 0.685 0.700 0.548 0.793 0.687

(c) The correlation coefficients between fibre weight and staple length in six cotton crosses were estimated as :

– 0.129, 0.1138, – 0.2780, 0.0033, 0.2331 and 0.0550

based on samples of sizes 73, 81, 67, 83, 71, 57 respectively. Test the homogeneity of  $r_i$ 's and obtain their best estimate.

**13-12. Non-central  $\chi^2$ -distribution.** The  $\chi^2$ -distribution defined as the sum of the squares of independent standard normal variates is often referred to as the *central  $\chi^2$ -distribution*. The distribution of the sum of the squares of independent normal variates each having unit variance but with possibly non-zero means is known as *non-central chi-square distribution*. Thus if  $X_i$ , ( $i = 1, 2, \dots, n$ ) are independent  $N(\mu_i, 1)$ , r.v.'s then

$$\chi^2 = \sum_{i=1}^n X_i^2, \quad \dots(13-24)$$

has the non-central  $\chi^2$  distribution with  $n$  d.f. Intuitively, this distribution would seem to depend upon the  $n$  parameters  $\mu_1, \mu_2, \dots, \mu_n$  but it will be seen that it depends on these parameters only through the *non-centrality parameter*

$$\lambda = \frac{1}{2}(\mu_1^2 + \mu_2^2 + \dots + \mu_n^2) \quad \dots(13.24a)$$

and we write  $\chi'^2 \sim \chi'^2(n, \lambda)$ .

**13.12.1. Non-central  $\chi^2$ -distribution with Non-centrality Parameter  $\lambda$ .** The p.d.f. is given by

$$f_{\chi'^2}(n, \lambda) = \sum_{j=0}^{\infty} \left[ \frac{e^{-\lambda} \lambda^j}{j!} \cdot p(\chi'^{2n+2j}) \right] \quad \dots(13.25)$$

where  $p(\chi'^{2n+2j})$  is the p.d.f. of (central)  $\chi^2$ -variate with  $n+2j$  d.f.

Thus  $f_{\chi'^2}(n, \lambda)$  is the mixture of central  $\chi^2$ -distributions with d.f.  $n, n+2, n+4, \dots$ , the corresponding weights being the successive terms of the Poisson distribution with parameter  $\lambda$ .

**Derivation of p.d.f. of  $\chi'^2$ .** We shall obtain the p.d.f. of non-central  $\chi^2$ -distribution through moment generating function (m.g.f.), by using the uniqueness theorem of m.g.f.

**13.12.2. Moment Generating Function of Non-central  $\chi^2$ -Distribution.** If  $X \sim N(\mu, 1)$  then

$$\begin{aligned} M_{X^2}(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} \cdot e^{-(x-\mu)^2/2} dx \\ \exp \left[ tx^2 - \frac{1}{2}(x-\mu)^2 \right] &= \exp \left[ - \left\{ \left( \frac{1}{2} - t \right) x^2 - \mu x + \frac{\mu^2}{2} \right\} \right] \\ &= \exp \left[ - \left( \frac{1-2t}{2} \right) \left\{ x^2 - \frac{2\mu x}{1-2t} + \frac{\mu^2}{1-2t} \right\} \right] \\ &= \exp \left[ - \left( \frac{1-2t}{2} \right) \left\{ \left( x - \frac{\mu}{1-2t} \right)^2 + \frac{\mu^2}{1-2t} - \frac{\mu^2}{(1-2t)^2} \right\} \right] \\ &= \exp \left( \frac{t\mu^2}{1-2t} \right) \exp \left[ - \left( \frac{1-2t}{2} \right) \left( x - \frac{\mu}{1-2t} \right)^2 \right] \\ \therefore M_{X^2}(t) &= \exp \left( \frac{t\mu^2}{1-2t} \right) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left[ - \left( \frac{1-2t}{2} \right) \left( x - \frac{\mu}{1-2t} \right)^2 \right] dx \\ &= \exp \left( \frac{t\mu^2}{1-2t} \right) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} u^2 \right) \frac{du}{(1-2t)^{1/2}} \\ &\doteq (1-2t)^{-1/2} \exp \left( \frac{t\mu^2}{1-2t} \right); 1-2t > 0 \Rightarrow t < \frac{1}{2} \quad \dots(*) \end{aligned}$$

If  $X_i$  ( $i = 1, 2, \dots, n$ ), are independent  $N(\mu_i, 1)$  then the m.g.f. of the non-central  $\chi^2$ -variate  $\chi'^2 = \sum_{i=1}^n X_i^2$  is given by

$$\begin{aligned}
 M_{\chi^2(n)}(t) &= M \sum_{i=1}^n X_i^2(t) = \prod_{i=1}^n M_{X_i^2}(t) \quad (\text{since } X_i^2 \text{ s are independent}) \\
 &= \prod_{i=1}^n \left[ (1-2t)^{-1/2} \exp \left( \frac{t\mu_i^2}{1-2t} \right) \right] \quad [\text{From (*)}] \\
 &= (1-2t)^{-n/2} \exp \left[ \frac{t}{(1-2t)} \sum_{i=1}^n \mu_i^2 \right] \\
 &= (1-2t)^{-n/2} \cdot \exp [2\lambda t / (1-2t)], \quad t < \frac{1}{2} \quad \dots(13-26)
 \end{aligned}$$

where  $\lambda = \frac{1}{2} \sum_{i=1}^n \mu_i^2$ , is the non-centrality parameter.

(13-26) can be re-written as:

$$\begin{aligned}
 M_{\chi^2(n)}(t) &= (1-2t)^{-n/2} \exp \left[ \lambda \left( -1 + \frac{1}{1-2t} \right) \right] \\
 &= (1-2t)^{-n/2} e^{-\lambda} \exp \left( \frac{\lambda}{1-2t} \right) \\
 &= (1-2t)^{-n/2} e^{-\lambda} \sum_{r=0}^{\infty} \left( \frac{\lambda}{1-2t} \right)^r \times \frac{1}{r!} \\
 &= \sum_{r=0}^{\infty} \frac{e^{-\lambda} \lambda^r}{r!} (1-2t)^{-(r+n/2)}; \quad t < \frac{1}{2} \quad \dots(13-26a)
 \end{aligned}$$

Thus the m.g.f. of a non-central  $\chi^2$  distribution is seen to be a *convex-combination* of  $\chi^2$  m.g.f.'s with d.f.  $n, n+2, n+4, \dots$ . The coefficients appearing in the convex combination are merely the Poisson probabilities.

Hence by the uniqueness theorem of m.g.f.'s the p.d.f. of non-central  $\chi^2$ -distribution with  $n$  d.f. and with non-centrality parameter  $\lambda$  is given by

$$f(\chi^2) = \sum_{r=0}^{\infty} \frac{e^{-\lambda} \lambda^r}{r!} \times p(\chi^2_{n+2r}),$$

$$\text{where } p(\chi^2_{n+2r}) = \frac{1}{2^{(n+2r)/2} \Gamma \left( \frac{n+2r}{2} \right)} e^{-\frac{1}{2}\chi^2} (\chi^2)^{\frac{n}{2}+r-1}; \quad 0 \leq \chi^2 < \infty$$

is the p.d.f. of central  $\chi^2$ -distribution with  $(n+2r)$  d.f.

**Remarks 1.** We can also write the m.g.f. of non-central  $\chi^2$  distribution with non-centrality parameter  $\lambda$  as

$$E \left[ ((1-2t))^{-\frac{n}{2}-Y} \right],$$

where  $Y$  is a Poisson variate with parameter  $\lambda$ .

2. If we take  $\lambda = 0 \Rightarrow \mu_i = 0 \forall i = 1, 2, \dots, n$ , the m.g.f. of the non-central  $\chi^2$  distribution reduces to the m.g.f. of central  $\chi^2$  distribution; viz.,  $(1-2t)^{-n/2}$ .

3. Taking  $\lambda = 0$  in the p.d.f. of non-central  $\chi^2$ -variate, i.e., in (13-25), we get

$$f(\chi'^2) = p(\chi_n^2) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-\frac{\lambda^2}{2}} (\chi^2)^{\frac{n}{2}-1}, \quad 0 \leq \chi^2 < \infty$$

[∴ we get contribution only when  $r = 0$ , the other terms vanish when  $\lambda = 0$ ]; which is p.d.f. of central  $\chi^2$ -distribution with  $n$  d.f.

**13-12-3. Additive or Re-productive Property of Non-central Chi-Square Distribution.** If  $Y_i$ , ( $i = 1, 2, \dots, k$ ), are independent non-central  $\chi^2$ -variates with  $n_i$  d.f. and non-centrality element  $\lambda_i$ , then  $\sum_{i=1}^k Y_i$  is also a non-central  $\chi^2$ -variante with  $\sum_{i=1}^k n_i$  d.f. and non-centrality element  $\lambda = \sum_{i=1}^k \lambda_i$ .

**Proof.** We have from (13-26),

$$M_{Y_i}(t) = (1 - 2t)^{-n_i/2} \exp [2t \lambda_i / (1 - 2t)], \quad (i = 1, 2, \dots, k)$$

$$\therefore M_{\sum Y_i}(t) = \prod_{i=1}^k M_{Y_i}(t) = (1 - 2t)^{-\sum n_i/2} \exp [2t \sum_i \lambda_i / (1 - 2t)],$$

which is the m.g.f. of a non-central  $\chi^2$ -variate with  $\sum n_i$  d.f. and non-centrality parameter  $\lambda = \sum \lambda_i$ . Hence by uniqueness theorem of m.g.f.'s

$$\sum_{i=1}^k Y_i \sim \chi'^2_{\sum n_i} (\sum \lambda_i)$$

**13-12-4. Cumulants of Non-central Chi-square Distribution.** Cumulant generating function is given by

$$\begin{aligned} K_{\chi'^2}(t) &= \log M_{\chi'^2}(t) = -\frac{n}{2} \log (1 - 2t) + 2t\lambda (1 - 2t)^{-1} \\ &= \frac{n}{2} \left[ 2t + \frac{(2t)^2}{2} + \dots + \frac{(2t)^r}{r} + \dots \right] + 2\lambda t \left[ 1 + 2t + (2t)^2 + \dots + (2t)^{r-1} + \dots \right] \end{aligned}$$

the expansion being valid for  $t < 1/2$ .

$$\begin{aligned} \therefore K_{\chi'^2}(t) &= (n + 2\lambda)t + (n + 4\lambda)t^2 + \dots + \left( \frac{2^{r-1}}{r} \cdot n + 2\lambda 2^{r-1} \right) t^r + \dots \\ \Rightarrow \quad \kappa_r &= \text{Coefficient of } \frac{t^r}{r!} \text{ in } K_{\chi'^2}(t) = r! \left( \frac{n}{r} + 2\lambda \right) 2^{r-1} \\ &\quad = 2^{r-1} (r-1)! (n+2\lambda r) \end{aligned} \quad \dots(13-27)$$

$$\therefore \kappa_{r-1} = 2^{r-2} (r-2)! [n + 2\lambda (r-1)]$$

$$\therefore \frac{d}{d\lambda} [\kappa_{r-1}] = 2^{r-2} (r-2)! 2(r-1) = 2^{r-1} (r-1)!$$

$$= \frac{\kappa_r}{(n + 2\lambda r)} \quad [\text{From (13-27)}]$$

$$\Rightarrow \kappa_r = (n + 2\lambda r) \frac{d}{d\lambda} (\kappa_{r-1}) \quad \dots(13-28)$$

# **Exact Sampling Distributions**

(CONTINUED)

(t, F AND Z DISTRIBUTIONS)

---

**14.1. Introduction.** The entire large sample theory was based on the application of "Normal Test" (c.f. § 12.9). However, if the sample size  $n$  is small, the distribution of the various statistics, e.g.,  $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$  or

$Z = (X - nP)/\sqrt{nPQ}$  etc., are far from normality and as such 'normal test' cannot be applied if  $n$  is small. In such cases exact sample tests, pioneered by W.S. Gosset (1908) who wrote under the pen name of Student, and later on developed and extended by Prof. R.A. Fisher (1926), are used. In the following sections we shall discuss

(i) t-test, (ii) F-test, and (iii) Fisher's z-transformation.

The exact sample tests can, however, be applied to large samples also though the converse is not true. In all the exact sample tests, the basic assumption is that "The population(s) from which sample(s) are drawn is (are) normal, i.e., the parent population(s) is (are) normally distributed."

**14.2. Student's 't'.** *Definition.* Let  $x_i$ , ( $i = 1, 2, \dots, n$ ) be a random sample of size  $n$  from a normal population with mean  $\mu$  and variance  $\sigma^2$ . Then Student's  $t$  is defined by the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \quad \dots(14.1)$$

where  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , is the sample mean and

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \dots(14.1a)$$

is an unbiased estimate of the population variance  $\sigma^2$ , and it follows Student's  $t$ -distribution with  $v = (n - 1)$  d.f. with probability density function,

$$f(t) = \frac{1}{\sqrt{v} B\left(\frac{1}{2}, \frac{v}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{v}\right]^{(v+1)/2}}; -\infty < t < \infty \quad \dots(14.2)$$

**Remarks 1.** A statistic  $t$  following Student's  $t$ -distribution with  $n$  d.f. will be abbreviated as  $t \sim t_n$ .

2. If we take  $v = 1$  in (14.2), we get

$$\begin{aligned} f(t) &= \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \cdot \frac{1}{(1+t^2)}, \\ &= \frac{1}{\pi} \cdot \frac{1}{(1+t^2)}; -\infty < t < \infty \quad [\because \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}] \end{aligned}$$

which is the p.d.f. of standard Cauchy distribution. Hence, when  $v = 1$  Student's  $t$ -distribution reduces to Cauchy distribution.

**14.2.1. Derivation of Student's  $t$ -distribution.** The expression (14.1) can be re-written as

$$\begin{aligned} t^2 &= \frac{n(\bar{x} - \mu)^2}{S^2} = \frac{n(\bar{x} - \mu)^2}{ns^2/(n-1)} \quad [\because ns^2 = (n-1) S^2] \\ \Rightarrow \frac{t^2}{(n-1)} &= \frac{(\bar{x} - \mu)^2}{\sigma^2/n} = \frac{1}{ns^2/\sigma^2} = \frac{(\bar{x} - \mu)^2/(\sigma^2/n)}{ns^2/\sigma^2} \end{aligned}$$

Since  $x_i$ , ( $i = 1, 2, \dots, n$ ) is a random sample from the normal population with mean  $\mu$  and variance  $\sigma^2$ ,

$$\bar{x} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Hence  $\frac{(\bar{x} - \mu)^2}{\sigma^2/n}$ , being the square of a standard normal variate is a chi-square variate with 1 d.f.

Also  $\frac{ns^2}{\sigma^2}$  is a  $\chi^2$ -variante with  $(n-1)$  d.f. (c.f. Theorem 13.5).

Further since  $\bar{x}$  and  $s^2$  are independently distributed (c.f. Theorem 13.5),  $\frac{t^2}{n-1}$ , being the ratio of two independent  $\chi^2$ -variates with 1 and  $(n-1)$  d.f. respectively, is a  $\beta_2\left(\frac{1}{2}, \frac{n-1}{2}\right)$  variante and its distribution is given by :

$$\begin{aligned} dF(t) &= \frac{1}{B\left(\frac{1}{2}, \frac{v}{2}\right)} \cdot \frac{(t^2/v)^{\frac{1}{2}-1}}{\left[1 + \frac{t^2}{v}\right]^{(v+1)/2}} dt; 0 \leq t^2 < \infty \\ &\quad [\text{where } v = (n-1)] \\ &= \frac{1}{\sqrt{v} B\left(\frac{1}{2}, \frac{v}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{v}\right]^{(v+1)/2}} dt; -\infty < t < \infty \end{aligned}$$

the factor 2 disappearing since the integral from  $-\infty$  to  $\infty$  must be unity. This is the required probability function as given in (14.2) of Student's  $t$ -distribution with  $v = (n-1)$  d.f.

**Remarks on Student's 't'.** 1. *Importance of Student's  $t$ -distribution in Statistics.* W.S. Gosset, who wrote under pseudonym (pen-name) of Student

defined his *t* in a slightly different way, viz.,  $t = (\bar{x} - \mu)/s$  and investigated its sampling distribution, somewhat empirically, in a paper entitled '*The probable error of the mean*', published in 1908. Prof. R.A. Fisher, later on defined his own '*t*' and gave a rigorous proof for its sampling distribution in 1926. The salient feature of '*t*' is that both the statistic and its sampling distribution are functionally independent of  $\sigma$ , the population standard deviation.

The discovery of '*t*' is regarded as a landmark in the history of statistical inference because of the following reason. Before Student gave his '*t*' it was customary to replace  $\sigma^2$  in  $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ , by its unbiased estimate  $S^2$  to give

$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$  and then normal test was applied even for small samples. It has been

found that although the distribution of *t* is asymptotically normal for large *n* (c.f. § 14-2-5), it is far from normality for small samples. The Student's *t* ushered in an era of exact sample distributions (and tests) and since its discovery many important contributions have been made towards the development and extension of small (exact) sample theory.

2. *Confidence or Fiducial Limits for  $\mu$* . If  $t_{0.05}$  is the tabulated value of *t* for  $v = (n - 1)$  d.f. at 5% level of significance, i.e.,

$$P(|t| > t_{0.05}) = 0.05 \Rightarrow P(|t| \leq t_{0.05}) = 0.95,$$

the 95% confidence limits for  $\mu$  are given by :

$$\begin{aligned} |t| &\leq t_{0.05}, \text{ i.e., } \left| \frac{\bar{x} - \mu}{S/\sqrt{n}} \right| \leq t_{0.05} \\ \Rightarrow \quad \bar{x} - t_{0.05} \cdot \frac{S}{\sqrt{n}} &\leq \mu \leq \bar{x} + t_{0.05} \cdot \frac{S}{\sqrt{n}} \end{aligned}$$

Thus, 95% confidence limits for  $\mu$  are :

$$\bar{x} \pm t_{0.05} \cdot \frac{S}{\sqrt{n}} \quad \dots [14.2(a)]$$

Similarly, 99% confidence limits for  $\mu$  are :

$$\bar{x} \pm t_{0.01} \cdot \frac{S}{\sqrt{n}} \quad \dots [14.2(b)]$$

where  $t_{0.01}$  is the tabulated value of *t* for  $v = (n - 1)$  d.f. at 1% level of significance.

14-2-2. *Fisher's 't'* (*Definition*). It is the ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom. If  $\xi$  is a  $N(0, 1)$  and  $\chi^2$  is an independent chi-square variate with *n* d.f., then Fisher's *t* is given by

$$t = \xi / \sqrt{\frac{\chi^2}{n}} \quad \dots (14.3)$$

and it follows student's '*t*' distribution with *n* degrees of freedom.

**14.2.3. Distribution of Fisher's 't'.** Since  $\xi$  and  $\chi^2$  are independent, their joint probability differential is given by

$$dF(\xi, \chi^2) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-\xi^2/2) \frac{\exp(-\chi^2/2) (\chi^2)^{\frac{n}{2}-1}}{2^{n/2} \Gamma(n/2)} d\xi d\chi^2$$

Let us transform to new variates  $t$  and  $u$  by the substitution

$$t = \frac{\xi}{\sqrt{\chi^2/n}} \text{ and } u = \chi^2 \Rightarrow \xi = t \sqrt{u/n} \text{ and } \chi^2 = u$$

Jacobian of transformation  $J$  is given by

$$J = \frac{\partial(\xi, \chi^2)}{\partial(t, u)} = \begin{vmatrix} \sqrt{u/n} & t/(2\sqrt{un}) \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{u}{n}}$$

The joint distribution of  $t$  and  $u$  becomes

$$dG(t, u) = \frac{1}{\sqrt{2\pi} 2^{n/2} \Gamma(n/2) \sqrt{n}} \exp\left\{-\frac{u}{2}\left(1 + \frac{t^2}{n}\right)\right\} u^{\frac{n}{2}-\frac{1}{2}} du dt;$$

Integrating w.r.t. 'u' over the range 0 to  $\infty$ , the marginal distribution of  $t$  becomes

$$\begin{aligned} dG_1(t) &= \frac{1}{\sqrt{2\pi} 2^{n/2} \Gamma(n/2) \sqrt{n}} \left[ \int_0^\infty \exp\left\{-\frac{u}{2}\left(1 + \frac{t^2}{n}\right)\right\} u^{(n-1)/2} du \right] dt \\ &= \frac{1}{\sqrt{2\pi} 2^{n/2} \Gamma(n/2) \sqrt{n}} \frac{\Gamma[(n+1)/2]}{\left[\frac{1}{2}\left(1 + \frac{t^2}{n}\right)\right]^{(n+1)/2}} dt \\ \therefore dG_1(t) &= \frac{\Gamma(n+1)/2}{\sqrt{n} \Gamma(n/2) \Gamma(\frac{1}{2})} \cdot \frac{1}{\left[1 + \frac{t^2}{n}\right]^{(n+1)/2}} dt, -\infty < t < \infty \\ &= \frac{1}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right) \left[1 + \frac{t^2}{n}\right]^{(n+1)/2}} dt, -\infty < t < \infty \end{aligned}$$

which is same as the probability function of Student's  $t$ -distribution with  $n$  d.f.

**Remarks 1.** In Fisher's ' $t$ ' the d.f. is the same as the d.f. of chi-square variate.

**2.** Student's ' $t$ ' may be regarded as a particular case of Fisher's ' $t$ ' as explained below.

$$\text{Since } \bar{x} \sim N(\mu, \sigma^2/n), \quad \xi = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \dots(*)$$

$$\text{and } \chi^2 = \frac{ns^2}{\sigma^2} = \sum_{i=1}^n (x_i - \bar{x})^2 / \sigma^2 \quad \dots(**)$$

is independently distributed as chi-square variate with  $(n - 1)$  d.f. Hence Fisher's *t* is given by

$$\begin{aligned} t &= \frac{\xi}{\sqrt{\chi^2/(n-1)}} = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \cdot \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2/(n-1)}} \\ &= \frac{\sqrt{n}(\bar{x} - \mu)}{S} = \frac{\bar{x} - \mu}{S/\sqrt{n}} \quad \dots (***) \end{aligned}$$

and it follows Student's *t*-distribution with  $(n - 1)$  d.f. (c.f. Remark 1 above.)

Now,  $(***)$  is same as Student's '*t*' defined in (14.1). Hence Student's '*t*' is a particular case of Fisher's '*t*'.

**14.2.4. Constants of *t*-distribution.** Since  $f(t)$  is symmetrical about the line  $t = 0$ , all the moments of odd order about origin vanish, i.e.,

$$\mu'_{2r+1} \text{ (about origin)} = 0 ; r = 0, 1, 2, \dots$$

In particular,

$$\mu'_1 \text{ (about origin)} = 0 = \text{Mean}$$

Hence central moments coincide with moments about origin.

$$\therefore \mu_{2r+1} = 0, (r = 1, 2, \dots) \quad \dots (14.4)$$

The moments of even order are given by

$$\mu_{2r} = \mu'_{2r} \text{ (about origin)}$$

$$\begin{aligned} &= \int_{-\infty}^{\infty} t^{2r} f(t) dt = 2 \int_0^{\infty} t^{2r} f(t) dt \\ &= 2 \cdot \frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)\sqrt{n}} \int_0^{\infty} \frac{t^{2r}}{\left[1 + \frac{t^2}{n}\right]^{(n+1)/2}} dt \end{aligned}$$

This integral is absolutely convergent if  $2r < n$ .

$$\text{Put } 1 + \frac{t^2}{n} = \frac{1}{y} \Rightarrow t^2 = n(1 - y)/y \text{ i.e., } 2tdt = -\frac{n}{y^2} dy$$

When  $t = 0, y = 1$  and when  $t = \infty, y = 0$ . Therefore,

$$\begin{aligned} \mu_{2r} &= \frac{2}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_1^0 \frac{t^{2r}}{(1/y)^{(n+1)/2}} \cdot \frac{-n}{2ty^2} dy \\ &= \frac{n}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 (t^2)^{(2r-1)/2} y^{[(n+1)/2]-2} dy \\ &= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 \left[n \left(\frac{1-y}{y}\right)\right]^{r-\frac{1}{2}} y^{[(n+1)/2]-2} dy \end{aligned}$$

$$\begin{aligned}
 &= \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 y^{\frac{n}{2}-r-1} (1-y)^{r-\frac{1}{2}} dy \\
 &= \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot B\left(\frac{n}{2}-r, r+\frac{1}{2}\right), n > 2r. \quad \dots [14.4(a)] \\
 &= n^r \frac{\Gamma[(n/2)-r] \Gamma(r + \frac{1}{2})}{\Gamma(\frac{1}{2}) \Gamma(n/2)} \\
 &= n^r \frac{(r - \frac{1}{2})(r - \frac{3}{2}) \dots \frac{3}{2} \frac{1}{2} \Gamma(\frac{1}{2}) \Gamma[(n/2) - r]}{\Gamma(\frac{1}{2}) [(n/2) - 1][(n/2) - 2] \dots [(n/2) - r] \Gamma[(n/2) - r]} \\
 &= n^r \frac{(2r-1)(2r-3)\dots 3 \cdot 1}{(n-2)(n-4)\dots(n-2r)} \cdot \frac{n}{2} > r \quad \dots [14.4(b)]
 \end{aligned}$$

In particular

$$\mu_2 = n \frac{1}{(n-2)} = \frac{n}{n-2}, [n > 2] \quad \dots [14.4(c)]$$

$$\text{and } \mu_4 = n^2 \frac{3 \cdot 1}{(n-2)(n-4)} = \frac{3n^2}{(n-2)(n-4)}, [n > 4] \quad \dots [14.4(d)]$$

$$\text{Hence } \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 \left( \frac{n-2}{n-4} \right)$$

**Remarks 1.** As  $n \rightarrow \infty$ ,  $\beta_1 = 0$  and

$$\beta_2 = \lim_{n \rightarrow \infty} 3 \left( \frac{n-2}{n-4} \right) = 3 \lim_{n \rightarrow \infty} \left[ \frac{1 - (2/n)}{1 - (4/n)} \right] = 3 \quad \dots [14.4(e)]$$

**2.** Changing  $r$  to  $(r-1)$  in [14.4(b)], dividing and simplifying, we shall get the recurrence relation for the moments as

$$\frac{\mu_{2r}}{\mu_{2r-2}} = \frac{n(2r-1)}{(n-2r)} \cdot \frac{n}{2} > r \quad \dots [14.4(f)]$$

**3. Moment Generating Function of t-distribution.** From [14.4(b)] we observe that if  $t \sim t_n$ , then all the moments of order  $2r < n$  exist but the moments of order  $2r \geq n$  do not exist. Hence the m.g.f. of t-distribution does not exist.

**Example 14.1.** Express the constants  $y_0$ ,  $a$  and  $m$  of the distribution :

$$dF(x) = y_0 \left[ 1 - \frac{x^2}{a^2} \right]^m dx, -a \leq x \leq a \quad \dots (*)$$

in terms of its  $\mu_2$  and  $\beta_2$ .

Show that if  $x$  is related to a variable  $t$  by the equation

$$x = \frac{at}{\sqrt{2(m+1) + t^2}}, \quad \dots (**)$$

then *t* has Student's distribution with  $2(m + 1)$  degrees of freedom. Use the transformation to calculate the probability that  $t \geq 2$  when the degrees of freedom are 2 and also when 4. (Madras Univ. M.Sc., 1991)

**Solution.** First of all we shall determine the constant from the consideration that total probability is unity.

$$\therefore y_0 \int_{-a}^a \left(1 - \frac{x^2}{a^2}\right)^m dx = 1$$

$$\Rightarrow 2y_0 \int_0^a \left(1 - \frac{x^2}{a^2}\right)^m dx = 1$$

( $\because$  Integrand is an even function of  $x$ )

$$\Rightarrow 2y_0 \int_0^{\pi/2} \cos^{2m} \theta \cdot a \cos \theta d\theta = 1 \quad (x = a \sin \theta)$$

$$\Rightarrow 2ay_0 \int_0^{\pi/2} \cos^{2m+1} \theta d\theta = 1$$

But we have the Beta integral,

$$2 \int_0^{\pi/2} \sin^p \theta \cos^q \theta d\theta = B\left(\frac{p+1}{2}, \frac{q+1}{2}\right) \quad \dots(1)$$

$$\therefore ay_0 \cdot 2 \int_0^{\pi/2} \cos^{2m+1} \theta \sin^0 \theta d\theta = 1$$

$$\Rightarrow ay_0 B(m+1, \frac{1}{2}) = 1 \quad [\text{Using (1)}]$$

$$\Rightarrow y_0 = \frac{1}{a B(m+1, \frac{1}{2})} \quad \dots(2)$$

Since the given probability function is symmetrical about the line  $x = 0$ , we have as in § 14.2-4.

$$\mu_{2r+1} = \mu_{2r'+1} = 0; r = 0, 1, 2, \dots \quad [\because \text{Mean} = \text{Origin}]$$

The moments of even order are given by

$$\mu_{2r} = \mu_{2r}' \text{ (about origin)}$$

$$= \int_{-a}^a x^{2r} f(x) dx = y_0 \int_{-a}^a x^{2r} \left(1 - \frac{x^2}{a^2}\right)^m dx$$

$$= 2y_0 \int_0^a x^{2r} \left(1 - \frac{x^2}{a^2}\right)^m dx$$

$$\begin{aligned}
 &= 2y_0 \int_0^{\pi/2} (a \sin \theta)^{2r} \cos^{2m} \theta \cdot a \cos \theta d\theta \quad [x = a \sin \theta] \\
 &= y_0 a^{2r+1} \cdot 2 \int_0^{\pi/2} \sin^{2r} \theta \cdot \cos^{2m+1} \theta d\theta \\
 &= y_0 a^{2r+1} B(r + \frac{1}{2}, m + 1) \quad [\text{Using (1)}] \\
 &= a^{2r} \frac{B(r + \frac{1}{2}, m + 1)}{B(m + 1, \frac{1}{2})} = a^{2r} \cdot \frac{\Gamma\left(r + \frac{1}{2}\right) \Gamma\left(m + \frac{3}{2}\right)}{\Gamma\left(m + r + \frac{3}{2}\right) \Gamma\left(\frac{1}{2}\right)} \dots (***) 
 \end{aligned}$$

In particular,  $\mu_2 = a^2 \cdot \frac{\Gamma(m + (3/2)) \cdot \frac{1}{2} \Gamma(1/2)}{(m + (3/2)) \Gamma(m + (3/2)) \Gamma(1/2)} = \frac{a^2}{2m + 3}$

$$\therefore a^2 = (2m + 3)\mu_2 \quad \dots (3)$$

Also  $\mu_4 = a^4 \frac{\Gamma(5/2)}{\Gamma(m + (7/2))} \times \frac{\Gamma(m + (3/2))}{\Gamma(1/2)}$   
 $= \frac{3a^4}{(2m + 5)(2m + 3)} \quad (\text{On simplification})$

 $\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3(2m + 3)}{(2m + 5)}$ 
 $\Rightarrow m = \frac{9 - 5\beta_2}{2(\beta_2 - 3)} \quad (\text{On simplification}) \dots (4)$

Equations (2), (3) and (4) express the constants  $y_0$ ,  $a$  and  $m$  in terms of  $\mu_2$  and  $\beta_2$ .

$$\begin{aligned}
 x &= \frac{at}{[2(m + 1) + t^2]^{1/2}} \Rightarrow \frac{x^2}{a^2} = \frac{t^2}{2(m + 1) + t^2} \\
 \text{i.e., } 1 - \frac{x^2}{a^2} &= \frac{2(m + 1)}{2(m + 1) + t^2} = \left(1 + \frac{t^2}{n}\right)^{-1}, \quad (n = 2m + 2) \\
 \text{Also } dx &= a \left[ \frac{dt}{(n + t^2)^{1/2}} - t \cdot \frac{1}{2} \frac{2t dt}{(n + t^2)^{3/2}} \right] \\
 &= a \frac{1}{(n + t^2)^{1/2}} \left[ 1 - \frac{t^2}{n + t^2} \right] dt \\
 &= \frac{an}{(n + t^2)^{3/2}} dt = \frac{a}{\sqrt{n}} \cdot \frac{1}{[1 + (t^2/n)]^{3/2}} dt
 \end{aligned}$$

Hence the p.d.f. of  $X$  transforms to

$$dF(t) = y_0 \frac{1}{\left[1 + \frac{t^2}{n}\right]^m} \cdot \frac{a}{\sqrt{n}} \frac{dt}{\left[1 + \frac{t^2}{n}\right]^{3/2}}$$

$$\begin{aligned}
 &= \frac{1}{a B(m+1, \frac{1}{2})} \cdot \frac{a}{\sqrt{n}} \frac{dt}{\left[1 + \frac{t^2}{n}\right]^{m+(3/2)}} \\
 &= \frac{1}{\sqrt{n} B(\frac{n}{2}, \frac{1}{2})} \cdot \frac{dt}{\left[1 + \frac{t^2}{n}\right]^{(n+1)/2}}, -\infty < t < \infty \quad \dots(5)
 \end{aligned}$$

which is the probability differential of Student's *t*-distribution with  $n = 2(m+1)$  d.f. Hence the result.

For 2 d.f. i.e.,  $n = 2$ , we get  $2(m+1) = 2 \Rightarrow m = 0$ . Hence from (\*\*), we get (for  $m = 0$ ),

$$x = \frac{at}{(2+t^2)^{1/2}} \Rightarrow x = \frac{\sqrt{2}}{\sqrt{3}} a, \text{ when } t = 2.$$

$$\begin{aligned}
 \therefore P(t \geq 2) &= P(X \geq \sqrt{(2/3)} a) = \int_{a\sqrt{(2/3)}}^a dF(x) \\
 &= \int_{a\sqrt{(2/3)}}^a \frac{1}{a B(1, \frac{1}{2})} dx \quad [\text{From (*), since } m = 0] \\
 &= \frac{1}{2a} \left( a - \frac{\sqrt{2}}{\sqrt{3}} a \right) = \frac{\sqrt{3} - \sqrt{2}}{2\sqrt{3}} \\
 &\quad \left[ \because B(1, \frac{1}{2}) = \frac{\Gamma(1)\Gamma(1/2)}{\Gamma(3/2)} = \frac{\Gamma(1/2)}{(1/2)\Gamma(1/2)} = 2 \right]
 \end{aligned}$$

For 4 d.f., i.e.,  $n = 4$ , we get  $m = 1$ . Proceeding exactly similarly we shall obtain

$$P(t \geq 2) = \frac{1}{2} - \frac{5\sqrt{2}}{16}$$

### EXERCISE 14(a)

1. (a) Given that

- (i)  $u$  is normally distributed with zero mean and unit variance,
- (ii)  $v^2$  has a chi-square distribution with  $n$  degrees of freedom, and
- (iii)  $u$  and  $v$  are independently distributed,

find the distribution of the variable

$$t = \frac{u\sqrt{n}}{v}$$

(b) Find the variance of the *t* distribution with  $n$  degrees of freedom, ( $n > 2$ ).

(c) If the variable *t* has Student's *t* distribution with 2 degrees of freedom, prove that

$$P(t \geq 2) = \frac{3 - \sqrt{6}}{6}$$

[Shivaji Univ. B.Sc., 1990]

2. (a) State, (without proof), the sampling distribution of Student's  $t$ . Who discovered it?

(b) 'Discovery of Student's  $t$  is regarded as a landmark in the history of statistical inference'. Elucidate.

(c) Let  $t$  be distributed as Student's  $t$ -distribution with 2 d.f. Find the probability  $P(-\sqrt{2} \leq t \leq \sqrt{2})$ .

3. (a) Show that

$$E(T^r) = \begin{cases} k^{r/2} \Gamma\left(\frac{1+r}{2}\right) \cdot \Gamma\left(\frac{k-r}{2}\right) \\ \frac{\Gamma(1/2) \cdot \Gamma(k/2)}{0, \text{ if } r \text{ is odd}} \end{cases}, \text{ if } r \text{ is even for } -1 < r < k$$

where  $T$  has Student's  $t$ -distribution with  $k$  degrees of freedom.

(b) For the  $t$ -distribution with  $n$  d.f., establish the recurrence relation

$$\mu_{2r} = \frac{n(2r-1)}{(n-2r)} \cdot \mu_{2r-2}, n > 2r$$

[Poona Univ. B.Sc., 1990; Delhi Univ. B.Sc. (Stat. Hons.), 1992]

(c) For how many d.f. does (i)  $\chi^2$ -distribution reduce to negative exponential distribution and (ii)  $t$ -distribution reduce to Cauchy distribution?

4. Suppose  $X_1, X_2, \dots, X_n$  ( $n > 1$ ) are independent variates each distributed as  $N(0, \sigma^2)$ . Find the p.d.f. of

$$W = X_1 \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2}^{1/2}$$

Why does not  $W$  follow the  $t$ -distribution?

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

5. Let  $x_1, x_2, \dots, x_n$  be independent observations from a normal universe with mean  $\mu$  and variance  $\sigma^2$  and let  $\bar{x}$  and  $s^2$  be the sample mean and sum of the squares of the deviations from the mean respectively. Let  $x'$  be one more observation independent of previous ones. Show that

$$\frac{x' - \bar{x}}{s} \left[ \frac{n(n-1)}{n+1} \right]^{1/2}$$

has a Student  $t$ -distribution with  $(n-1)$  degrees of freedom.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

6. (a) Let  $X_1$  and  $X_2$  be two independent normal variates with the same normal distribution  $N(\mu, \sigma^2)$ . Obtain the distribution of

$$Y = \frac{X_1 + X_2 - 2\mu}{\sqrt{|X_1 - X_2|^2}}$$

Ans. Standard Cauchy distribution.

(b) If  $X$  is *t*-distributed with  $k$  degrees of freedom, show that

$$\frac{1}{1 + (X^2/k)},$$

has a beta distribution.

[Delhi Univ. B.Sc. (Maths. Hons.), 1988]

7. Define Student's *t*-statistic and state its probability density function.

If  $x_i$  ( $i = 1, 2, \dots, n$ ), is a random sample of  $n$  independent observations from a normal population with mean  $\mu$  and variance  $\sigma^2$ , show that

$$U = \frac{(\bar{x} - \mu) \sqrt{n(n-1)}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, \text{ where } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

conforms to Student's *t*-variate. If  $x$  is an additional observation drawn independently from the same normal population, show that

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

also conforms to Student's *t*-variate.

8. Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$ , and  $\bar{X}$  and  $S^2$ , respectively, be the sample mean and sample variance. Let  $X_{n+1} \sim N(\mu, \sigma^2)$ , and assume that  $X_1, X_2, \dots, X_n, X_{n+1}$  are independent. Obtain the sampling distribution of

$$U = \frac{(X_{n+1} - \bar{X})}{S} \cdot \sqrt{\frac{n}{n+1}}; \quad \left[ S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right]$$

9. If the random variables  $X_1$  and  $X_2$  are independent and follow chi-square distribution with  $n$  d.f., show that  $\frac{\sqrt{n}(X_1 - \bar{X}_2)}{2\sqrt{X_1 X_2}}$  is distributed as Student's *t* with  $n$  d.f., independently of  $X_1 + X_2$ .

[Calcutta Univ. B.Sc. (Hons.), 1992]

**Hint.**  $p(x_1, x_2) = \frac{1}{2^n [\Gamma(n/2)]^2} \cdot e^{-(x_1 + x_2)/2} x_1^{(n/2)-1} x_2^{(n/2)-1};$

$$0 \leq x_1 < \infty, 0 \leq x_2 < \infty$$

Put  $u = \frac{\sqrt{n}(x_1 - x_2)}{2\sqrt{x_1 x_2}}$  and  $v = x_1 + x_2$

$$\Rightarrow x_1 = \frac{v}{2} \left[ 1 + \frac{1}{\sqrt{\left(1 + \frac{n}{u^2}\right)}} \right], \quad x_2 = \frac{v}{2} \left[ 1 - \frac{1}{\sqrt{\left(1 + \frac{n}{u^2}\right)}} \right]$$

Jacobian of transformation is  $J = \frac{\partial(x_1, x_2)}{\partial(u, v)} = \frac{v}{2\sqrt{n} [1 + u^2/n]^{3/2}}$

The joint p.d.f. of  $U$  and  $V$  becomes

$$g(u, v) = p(x_1, x_2) |J| = \frac{1}{2^{2n-1} \Gamma(n/2) \Gamma(n/2) \sqrt{n}} \cdot \frac{e^{-v/2} v^{n-1}}{(1 + u^2/n)^{(n+1)/2}}; \quad -\infty < u < \infty, 0 \leq v < \infty$$

Using Legendre's duplication formula, viz.,

$$\Gamma n = 2^{n-1} \Gamma(n/2) \Gamma\left(\frac{n+1}{2}\right) \sqrt{\pi} \Rightarrow \Gamma(n/2) = \frac{\Gamma n \sqrt{\pi}}{2^{n-1} \Gamma\left(\frac{n+1}{2}\right)}, \text{ we get}$$

$$\begin{aligned} 2^{2n-1} \Gamma(n/2) \Gamma(n/2) \sqrt{n} &= \frac{2^{2n-1} \sqrt{n} \sqrt{\pi}}{2^{n-1} \Gamma\left(\frac{n+1}{2}\right)} \Gamma\left(\frac{n}{2}\right) \sqrt{n} \\ &= 2^n \sqrt{n} \sqrt{n} B\left(\frac{1}{2}, n/2\right) \quad [\because \sqrt{\pi} = \Gamma\left(\frac{1}{2}\right)] \end{aligned}$$

$$g(u, v) = \left( \frac{1}{2^n \Gamma n} e^{-v/2} v^{n-1} \right) \left[ \frac{1}{\sqrt{n} B\left(\frac{1}{2}, n/2\right)} \cdot \frac{1}{\left(1 + \frac{u^2}{n}\right)^{(n+1)/2}} \right];$$

$$0 < v < \infty, -\infty < u < \infty.$$

10. Let  $X_1, X_2, \dots, X_m$  and  $Y_1, Y_2, \dots, Y_n$  be independent random samples from  $N(\mu_1, \sigma^2)$  and  $N(\mu_2, \sigma^2)$ , respectively. If  $\bar{X}$  and  $\bar{Y}$  denote the corresponding sample means and if

$$(m-1)S_1^2 = \sum_{i=1}^m (X_i - \bar{X})^2, \quad (n-1)S_2^2 = \sum_{j=1}^n (Y_j - \bar{Y})^2,$$

obtain the sampling distribution of

$$\frac{a(\bar{X} - \mu_1) + b(\bar{Y} - \mu_2)}{\left[ \frac{(m-1)S_1^2 + (n-1)S_2^2}{(m+n-2)} \right] \left\{ \frac{a^2}{m} + \frac{b^2}{n} \right\}}^{1/2}$$

where  $a$  and  $b$  are two fixed real numbers.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

11. If  $I_x(p, q)$  represents the incomplete Beta function defined by

$$I_x(p, q) = \frac{1}{B(p, q)} \int_0^x t^{p-1} (1-t)^{q-1} dt; \quad p > 0, q > 0,$$

show that the distribution function  $F(\cdot)$  of Student's  $t$ -distribution is given by

$$F(t) = 1 - \frac{1}{2} I_x\left(\frac{n}{2}, \frac{1}{2}\right), \text{ where } x = \left(1 + \frac{t^2}{n}\right)^{-1}.$$

[Delhi Univ. M.Sc. (Stat.), 1990; Nagpur Univ. M.Sc. (Stat.), 1991]

**Hint.** If  $f(t)$  is p.d.f. of *t*-distribution with  $n$  d.f., then

$$\begin{aligned}
 F(t) &= \int_{-\infty}^t f(u) du = 1 - \int_t^{\infty} f(u) du \\
 &= 1 - \frac{1}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_t^{\infty} \left(1 + \frac{u^2}{n}\right)^{-(n+1)/2} du \\
 &= 1 + \frac{1}{2 B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^{\infty} \left(1 + \frac{t^2}{n}\right)^{-1} z^{(n/2)-1} (1-z)^{-1/2} dz, \\
 &\quad \text{where } \left[\frac{1}{z} = 1 + \frac{u^2}{n}\right] \\
 &= 1 - \frac{1}{2 B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^x z^{(n/2)-1} (1-z)^{-1/2} dz, \quad \left[x = \left(1 + \frac{t^2}{n}\right)^{-1}\right] \\
 &= 1 - \frac{1}{2} I_x\left(\frac{n}{2}, \frac{1}{2}\right)
 \end{aligned}$$

12. Show that for *t*-distribution with  $n$  d.f., mean deviation about mean is given by

$$\sqrt{n} \Gamma\left(\frac{n-1}{2}\right) / \sqrt{\pi} \Gamma(n/2)$$

(Shivaji Univ. B.Sc. Oct., 1992]

**Hint.**  $E(t) = 0$ .

$$\begin{aligned}
 \text{M.D. about mean} &= \int_{-\infty}^{\infty} |t| f(t) dt \\
 &= \frac{1}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{-\infty}^{\infty} \frac{|t| dt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}} \\
 &= \frac{2}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^{\infty} \frac{tdt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}} \\
 &= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^{\infty} \frac{dy}{(1+y)^{(n+1)/2}}, \quad \left[\left(\frac{t^2}{n} = y\right)\right]
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^\infty \frac{y^{1-1}}{(1+y)^{\frac{n-1}{2}+1}} dy \\
 &= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} B\left(\frac{n-1}{2}, 1\right)
 \end{aligned}$$

13. If  $X \sim t_{(n)}$ , show that

$$(n - \frac{1}{2}) \log \left[ 1 + \frac{x^2}{n} \right] \sim \chi^2_{(1)},$$

for large  $n$ .

You may assume that for large  $n$ ,

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

14. If  $\bar{X}$  and  $\hat{\sigma}^2 = S^2$  be the usual sample mean and sample variance based on a random sample of  $n$  observations from  $N(\mu, \sigma^2)$ , and if  $T = (\bar{X} - \mu) / \sqrt{n}/S$ , prove that

$$(i) \text{Var}(T) = (n-1)/(n-3)$$

$$(ii) \text{Cov}(\bar{X}, T) = \sigma \frac{\sqrt{n-1}}{\sqrt{2n}} \frac{\Gamma[(n-2)/2]}{\Gamma[(n-1)/2]}$$

$$(iii) r(\bar{X}, T) = [\frac{1}{2}(n-3)]^{1/2} \frac{\Gamma[\frac{1}{2}(n-2)]}{\Gamma[\frac{1}{2}(n-1)]}$$

**14.2.5. Limiting Form of t-distribution.** As  $n \rightarrow \infty$ , the p.d.f. of t-distribution with  $n$  d.f. viz.,

$$f(t) = \frac{1}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \rightarrow \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad -\infty < t < \infty$$

$$\begin{aligned}
 \text{Proof. } \lim_{n \rightarrow \infty} \frac{1}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)} &= \lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} \frac{\Gamma[(n+1)/2]}{\Gamma(\frac{1}{2}) \Gamma(n/2)} \\
 &= \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\pi}} \left(\frac{n}{2}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2\pi}}
 \end{aligned}$$

$$\left[ \because \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \text{ and } \lim_{n \rightarrow \infty} \frac{\Gamma(n+k)}{\Gamma(n)} = n^k, \text{ (c.f. Remark to § 14.5.7)} \right]$$

$$\begin{aligned} \therefore \lim_{n \rightarrow \infty} f(t) &= \lim_{n \rightarrow \infty} \frac{1}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \lim_{n \rightarrow \infty} \left[ \left(1 + \frac{t^2}{n}\right)^n \right]^{-\frac{1}{2}} \\ &\quad \times \lim_{n \rightarrow \infty} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}} \\ &= \frac{1}{\sqrt{2\pi}} \exp(-t^2/2), \quad -\infty < t < \infty \end{aligned}$$

Hence for large d.f. *t*-distribution tends to standard normal distribution.

**14.2.6. Graph of *t*-distribution.** The p.d.f. of *t*-distribution with *n* d.f. is

$$f(t) = C \cdot \left[1 + \frac{t^2}{n}\right]^{-(n+1)/2}, \quad -\infty < t < \infty$$

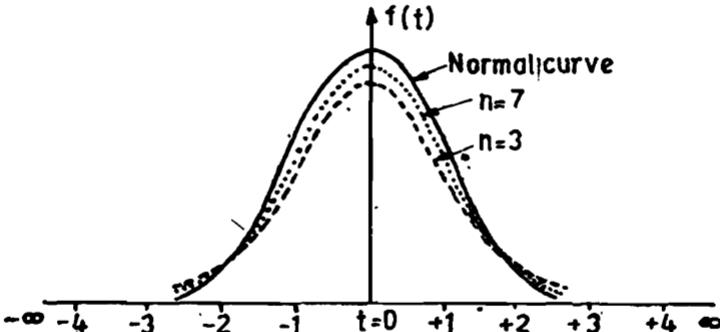
Since  $f(-t) = f(t)$ , the probability curve is symmetrical about the line  $t = 0$ . As  $t$  increases,  $f(t)$  decreases rapidly and tends to zero as  $t \rightarrow \infty$ , so that  $t$ -axis is an asymptote to the curve. We have shown that

$$\mu_2 = \frac{n}{n-2}, \quad n > 2; \quad \beta_2 = \frac{3(n-2)}{(n-4)}, \quad n > 4$$

Hence for  $n > 2$ ,  $\mu_2 > 1$  i.e., the variance of *t*-distribution is greater than that of standard normal distribution and for  $n > 4$ ,  $\beta_2 > 3$  and thus *t*-distribution is more flat on the top than the normal curve. In fact, for small *n*, we have

$$P[|t| \geq t_0] \geq P[|Z| \geq t_0], \quad Z \sim N(0, 1)$$

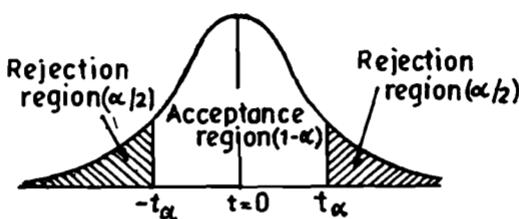
i.e., the tails of the *t*-distribution have a greater probability (area) than the tails of standard normal distribution. Moreover we have also seen [§ 14.2.5] that for large *n* (d.f.), *t*-distribution tends to standard normal distribution.



**14.2.7. Critical Values of *t*.** The critical (or significant) values of *t* at level of significance  $\alpha$  and d.f. *v* for two-tailed test are given by the equation:

$$P[|t| > t_v(\alpha)] = \alpha \quad \dots(14.5)$$

$$\Rightarrow P[|t| \leq t_v(\alpha)] = 1 - \alpha. \quad \dots(14.5a)$$

CRITICAL VALUES OF  $t$ -DISTRIBUTION

The values  $t_v(\alpha)$  have been tabulated in Fisher and Yates' Tables, for different values of  $\alpha$  and  $v$  and are given in the Appendix at the end of the book.

Since  $t$ -distribution is symmetric about  $t = 0$ , we get from (14.5)

$$\begin{aligned} P[t > t_v(\alpha)] + P[t < -t_v(\alpha)] &= \alpha \\ \Rightarrow 2P[t > t_v(\alpha)] &= \alpha \\ \Rightarrow P[t > t_v(\alpha)] &= \alpha/2 \\ \Rightarrow P[t > t_v(2\alpha)] &= \alpha \end{aligned} \quad \dots(14.5b)$$

$t_v(2\alpha)$  (from the Tables in the Appendix) gives the significant value of  $t$  for a single-tail test, [Right-tail or Left-tail-since the distribution is symmetrical], at level of significance  $\alpha$  and  $v$  d.f.

Hence the significant values of  $t$  at level of significance ' $\alpha$ ' for a single tailed test can be obtained from those of two-tailed test by looking the values at level of significance ' $2\alpha$ '.

For example,

$t_8(0.05)$  for single-tail test =  $t_8(0.10)$  for two-tail test = 1.86

$t_{15}(0.01)$  for single-tail test =  $t_{15}(0.02)$  for two-tail test = 2.60.

**14.2.8. Applications of  $t$ -distribution.** The  $t$ -distribution has a wide number of applications in Statistics, some of which are enumerated below.

(i) To test if the sample mean ( $\bar{x}$ ) differs significantly from the hypothetical value  $\mu$  of the population mean.

(ii) To test the significance of the difference between two sample means.

(iii) To test the significance of an observed sample correlation co-efficient and sample regression coefficient.

(iv) To test the significance of observed partial and multiple correlation coefficients.

In the following sections we will discuss these applications in detail, one by one.

**14.2.9.  $t$ -Test for Single Mean.** Suppose we want to test :

(i) if a random sample  $x_i$  ( $i = 1, 2, \dots, n$ ) of size  $n$  has been drawn from a normal population with a specified mean, say  $\mu_0$ , or

(ii) if the sample mean differs significantly from the hypothetical value  $\mu_0$  of the population mean.

Under the null hypothesis  $H_0$ :

(i) The sample has been drawn from the population with mean  $\mu$  or (ii) there is no significant difference between the sample mean  $\bar{x}$  and the population mean  $\mu$ ,

the statistic

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \quad \dots(14.6)$$

where  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  and  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ ,  $\dots[14.6(a)]$

follows Student's *t*-distribution with  $(n-1)$  d.f.

We now compare the calculated value of  $t$  with the tabulated value at certain level of significance. If calculated  $|t| >$  tabulated  $t$ , null hypothesis is rejected and if calculated  $|t| <$  tabulated  $t$ ,  $H_0$  may be accepted at the level of significance adopted.

**Remarks 1.** On computation of  $S^2$  for numerical problems. If  $\bar{x}$  comes out in integers, the formula (14.6a) can be conveniently used for computing  $S^2$ . However, if  $\bar{x}$  comes in fractions then the formula (14.6a) for computing  $S^2$  is very cumbersome and is not recommended. In that case, step deviation method, given below, is quite useful.

If we take,  $d_i = x_i - A$ , where  $A$  is any arbitrary number then

$$S^2 = \frac{1}{n-1} \left[ \sum (x_i - \bar{x})^2 \right] = \frac{1}{n-1} \left[ \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] \quad \dots[14.6(b)]$$

$$= \frac{1}{n-1} \left[ \sum d_i^2 - \frac{(\sum d_i)^2}{n} \right], \quad \dots[14.6(c)]$$

since variance is independent of change of origin.

$$\text{Also, in this case } \bar{x} = A + \frac{\sum d_i}{n}. \quad \dots[14.6(d)]$$

2. We know, the sample variance

$$\begin{aligned} S^2 &= \frac{1}{n} \sum_i (x_i - \bar{x})^2 \\ ns^2 &= (n-1) S^2 \\ \Rightarrow \quad \frac{S^2}{n} &= \frac{s^2}{n-1} \end{aligned} \quad \dots[14.6(e)]$$

Hence for numerical problems, the test statistic (14.6) on using [14.6(e)] becomes

$$t = \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} = \frac{\bar{x} - \mu_0}{\sqrt{s^2/(n-1)}} \sim t_{n-1} \quad \dots[14.6(f)]$$

3. Assumptions for Student's *t*-test. The following assumptions are made in the Student's *t*-test :

- (i) The parent population from which the sample is drawn is normal.
- (ii) The sample observations are independent, i.e., the sample is random.
- (iii) The population standard deviation  $\sigma$  is unknown.

**Example 14-2.** A machinist is making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a standard deviation of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.

**Solution.** Here we are given :

$$\mu = 0.700 \text{ inches}, \bar{x} = 0.742 \text{ inches}, s = 0.040 \text{ inches and } n = 10$$

**Null Hypothesis.**  $H_0 : \mu = 0.700$ , i.e., the product is conforming to specifications.

**Alternative Hypothesis.**  $H_1 : \mu \neq 0.700$

**Test Statistic.** Under  $H_0$ , the test statistic is :

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} = \frac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{(n-1)}$$

$$\text{Now } t = \frac{\sqrt{9(0.742 - 0.700)}}{0.040} = 3.15$$

**How to proceed further.** Here the test statistic ' $t$ ' follows Student's  $t$ -distribution with  $10 - 1 = 9$  d.f. We will now compare this calculated value with the tabulated value of  $t$  for 9 d.f. and at certain level of significance, say 5%. Let this tabulated value be denoted by  $t_0$ .

(i) If calculated ' $t$ ' viz.,  $3.15 > t_0$ , we say that the value of  $t$  is significant. This implies that  $\bar{x}$  differs significantly from  $\mu$  and  $H_0$  is rejected at this level of significance and we conclude that the product is not meeting the specifications.

(ii) If calculated  $t < t_0$ , we say that the value of  $t$  is not significant, i.e., there is no significant difference between  $\bar{x}$  and  $\mu$ . In other words, the deviation ( $\bar{x} - \mu$ ) is just due to fluctuations of sampling and null hypothesis  $H_0$  may be retained at 5% level of significance, i.e., we may take the product conforming to specifications.

**Example 14-3.** 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.** We are given :  $n = 22$ ,  $\bar{x} = 153.7$ ,  $s = 17.2$ .

**Null Hypothesis.** The advertising campaign is not successful, i.e.,

$$H_0 : \mu = 146.3$$

**Alternative Hypothesis.**  $H_1 : \mu > 146.3$  (Right-tail).

**Test Statistic.** Under the null hypothesis, the test statistic is :

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{22-1} = t_{21}$$

$$\text{Now } t = \frac{153.7 - 146.3}{\sqrt{(17.2)^2/21}} = \frac{7.4 \times \sqrt{21}}{17.2} = 9.03$$

**Conclusion.** Tabulated value of  $t$  for 21 d.f. at 5% level of significance for single-tailed test is 1.72. Since calculated value is much greater than the

tabulated value, it is highly significant. Hence we reject the null hypothesis and conclude that the advertising campaign was definitely successful in promoting sales.

**Example 14-4.** A random sample of 10 boys had the following I.Q.'s : 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100 ? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.

[*Madras Univ. B.E., April 1990*]

**Solution.** Null hypothesis,  $H_0$  : The data are consistent with the assumption of a mean I.Q. of 100 in the population, i.e.,  $\mu = 100$ .

Alternative hypothesis,  $H_1$  :  $\mu \neq 100$ .

**Test Statistic.** Under  $H_0$ , the test statistic is :

$$t = \frac{(\bar{x} - \mu)}{\sqrt{S^2/n}} \sim t_{(n-1)}$$

where  $\bar{x}$  and  $S^2$  are to be computed from the sample values of I.Q.'s.

#### CALCULATIONS FOR SAMPLE MEAN AND S.D.

X	(X - $\bar{x}$ )	(X - $\bar{x}$ ) <sup>2</sup>
70	-27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
Total 972		1833.60

Hence  $n = 10$ ,  $\bar{x} = \frac{972}{10} = 97.2$  and  $S^2 = \frac{1833.60}{9} = 203.73$

$$\therefore |t| = \frac{|97.2 - 100|}{\sqrt{203.73/10}} = \frac{2.8}{\sqrt{20.37}} = \frac{2.8}{4.514} = 0.62$$

Tabulated  $t_{0.05}$  for  $(10 - 1)$  i.e., 9 d.f. for two-tailed test is 2.262.

**Conclusion.** Since calculated  $t$  is less than tabulated  $t_{0.05}$  for 9 d.f.,  $H_0$  may be accepted at 5% level of significance and we may conclude that the data are consistent with the assumption of mean I.Q. of 100 in the population.

The 95% confidence limits within which the mean I.Q. values of samples of 10 boys will lie are given by

$$\bar{x} \pm t_{0.05} S / \sqrt{n} = 97.2 \pm 2.262 \times 4.514$$

$$= 97.2 \pm 10.21 = 107.41 \text{ and } 86.99$$

Hence the required 95% confidence interval is [86.99, 107.41].

**Remark.** Aliter for computing  $\bar{x}$  and  $S^2$ . Here we see that  $\bar{x}$  comes in fractions and as such the computation of  $(x - \bar{x})^2$  is quite laborious and time consuming. In this case we use the method of step deviations to compute  $\bar{x}$  and  $S^2$ , as given below.

X	d = X - 90	$d^2$
70	-20	400
120	30	900
110	20	400
101	11	121
88	-2	4
83	-7	49
95	5	25
98	8	64
107	17	289
100	10	100
Total	$\sum d = 72$	$\sum d^2 = 2352$

Here  $d = X - A$ , where  $A = 90$

$$\therefore \bar{x} = A + \frac{1}{n} \sum d = 90 + \frac{72}{10} = 97.2$$

$$\text{and } S^2 = \frac{1}{n-1} \left[ \sum d^2 - \frac{(\sum d)^2}{n} \right] = \frac{1}{9} \left[ 2352 - \frac{(72)^2}{10} \right] = 203.73$$

**Example 14-5.** The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% significance level, assuming that for 9 degrees of freedom  $P(t > 1.83) = 0.05$ .

**Solution.** Null Hypothesis,  $H_0 : \mu = 64$  inches.

Alternative Hypothesis,  $H_1 : \mu > 64$  inches.

#### CALCULATIONS FOR SAMPLE MEAN AND S.D.

x	70	67	62	68	61	68	70	64	64	66	Total 660
$x - \bar{x}$	4	1	-4	2	-5	2	4	-2	-2	0	0
$(x - \bar{x})^2$	16	1	16	4	25	4	16	4	4	0	90

$$\bar{x} = \frac{\sum x}{n} = \frac{660}{10} = 66$$

$$S^2 = \frac{1}{n-1} \sum (x - \bar{x})^2 = \frac{90}{9} = 10$$

*Test Statistic.* Under  $H_0$ , the test statistic is

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{66 - 64}{\sqrt{10/10}} = 2,$$

which follows Student's *t*-distribution with  $10 - 1 = 9$  d.f.

Tabulated value of *t* for 9 d.f. at 5% level of significance for single (right) tail-test is 1.833. (This is the value  $t_{0.10}$  for 9 d.f. in the two-tailed Table given in the Appendix.)

*Conclusion.* Since calculated value of *t* is greater than the tabulated value, it is significant. Hence  $H_0$  is rejected at 5% level of significance and we conclude that the average height is greater than 60 inches.

**Example 14.6.** 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. Obtain 95 per cent and 99 per cent fiducia. limits for the same.

You may use the following information from statistical tables :

$$v = 15, \begin{cases} P = 0.05, t = 2.131 \\ P = 0.01, t = 2.947 \end{cases}$$

**Solution.** We are given  $n = 16$ ,  $\bar{x} = 41.5$  inches and

$$\sum (x - \bar{x})^2 = 135 \text{ sq. inches.}$$

$$\therefore S^2 = \frac{1}{n-1} \sum (x - \bar{x})^2 = \frac{135}{15} = 9 \Rightarrow S = 3.$$

*Null Hypothesis.*  $H_0 : \mu = 43.5$  inches, i.e., the data are consistent with the assumption that the mean height in the population is 43.5 inches.

*Alternative Hypothesis.*  $H_1 : \mu \neq 43.5$  inches.

*Test Statistic.* Under  $H_0$ , the test statistic is :

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

$$\text{Now } |t| = \frac{|41.5 - 43.5|}{3/4} = \frac{8}{3} = 2.667$$

Here number of degrees of freedom is  $(16 - 1) = 15$ .

We are given :

$$t_{0.05} \text{ for 15 d.f.} = 2.131 \text{ and } t_{0.01} \text{ for 15 d.f.} = 2.947$$

*Conclusion.* Since calculated  $|t|$  is greater than 2.131, null hypothesis is rejected at 5% level of significance and we conclude that the assumption of mean of 43.5 inches for the population is not reasonable.

**Remark.** Since calculated  $|t|$  is less than 2.947, null hypothesis ( $\mu = 43.5$ ) may be accepted at 1% level of significance.

**95% fiducial limits for  $\mu$  : (d.f. = 15)**

$$\bar{x} \pm t_{0.05} \times \frac{S}{\sqrt{n}} = 41.5 \pm 2.131 \times \frac{3}{4} = 41.5 \pm 1.598$$

$$\therefore 39.902 < \mu < 43.098$$

**99% fiducial limits for  $\mu$  : (d.f. = 15)**

$$\bar{x} \pm t_{0.01} \times \frac{S}{\sqrt{n}} = 41.5 \pm 2.947 \times \frac{3}{4} = 43.71 \text{ and } 39.29$$

$$\therefore 39.29 < \mu < 43.71$$

### EXERCISE 14(b)

1. (a) Write a short note on Student's  $t$ -distribution and point out its uses.

(b) Show how the  $t$ -distribution has been found useful in testing whether the mean of small sample is significantly different from a hypothetical value.

(c) It is desired to test the hypothesis that the mean of a normal population is  $\mu = \mu_0$  against the alternative that  $\mu \neq \mu_0$ . Explaining the assumptions involved, develop the statistic suitable for testing this hypothesis if the size of the sample is small. What modification do you suggest when the sample size is large?

2. What is a test of significance ?

To test the hypothesis that the mean of a normal distribution is zero, two independent observations  $x_1$  and  $x_2$  are taken from the distribution. Show that the hypothesis is rejected at 10% level of significance, using  $t$  test with equal tail ends, if

$$|x_1 + x_2| > |x_1 - x_2| \tan 81^\circ$$

3. It is required to test that the mean of a normal population is zero. A random sample drawn from the population gives the values  $x_1, x_2, \dots, x_n$ . Show that the  $t$ -test for acceptance of the hypothesis reduces to

$$\left( \sum_{i=1}^n x_i \right) \leq \frac{n \cdot t_\alpha^2}{t_\alpha^2 + (n-1)} \left( \sum_{i=1}^n x_i^2 \right)$$

where  $t_\alpha$  is the value of Student's  $t$  at the desired level of significance  $\alpha$  for  $(n-1)$  d.f.

4. (a) Find the Student's  $t$  for following variate values in a sample of eight : -4, -2, -2, 0, 2, 2, 3, 3, taking the mean of the universe to be zero. How would you proceed further ?

(b) Ten individuals are chosen at random from a normal population and their heights are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71 inches. Test if the sample belongs to the population whose mean height is 66"

[Given  $t_{0.05} = 2.62$  for 9 d.f.]

(c) A random sample of 9 experimental animals under a certain diet gave the following increase in weight :  $\sum x_i = 45$  lbs,  $\sum x_i^2 = 279$  lbs., where  $x_i$  denotes the increase in weight of the  $i$ th animal. Assuming that the increase in weight is normally distributed as  $N(\mu, \sigma^2)$  variate, test  $H_0 : \mu = 1$  against  $H_1 : \mu \neq 1$  at 5% level. Given  $P(|t| > 2.306) = 0.05$  for 8 degrees of freedom.

[Calcutta Univ. B.Sc. (Maths.Hons.), 1991]

5. A manufacturer of gunpowder has developed a new powder which is designed to produce a muzzle velocity equal to 3000 ft/sec. Seven shells are loaded with the charge and the muzzle velocities measured.

The resulting velocities are as follows : 3,005; 2,935; 2,965; 2,995; 3,905, 2,935; and 2,905. Do these data present sufficient evidence to indicate that the average velocity differs from 3,000 ft/sec.

6. The average length of time for students to register for summer classes at a certain college has been 50 minutes with a standard deviation of 10 minutes. A new registration procedure using modern computing machines is being tried. If a random sample of 12 students had an average registration time of 42 minutes with s.d. of 11.9 minutes under the new system, test the hypothesis that the population mean has not changed, using .05 as level of significance.

7. The nine items of a sample had the following values : 45, 47, 50, 52, 48, 47, 49, 53 and 51.

Does the mean of the nine items differ significantly from the assumed population mean of 47.5 ? Given that

$$v = 8, \begin{cases} P = 0.945 \text{ for } t = 1.8 \\ P = 0.953 \text{ for } t = 1.9 \end{cases}$$

8. A time study engineer developed a new sequence of operation elements that he hopes will reduce the mean cycle time of a certain production process. The results of a time study of 20 cycles are given below :

*cycle time in minutes*

12.25	11.97	12.15	12.08	12.31	12.28	11.94	11.89	12.16	12.04
12.09	12.15	12.14	12.47	11.98	12.04	12.11	12.25	12.15	12.34

If the present mean cycle time is 12.5 minutes, should he adopt the new sequence ?

9. (a) The average breaking strength of steel rods is specified to be 18.5 thousand pounds. To test this a sample of 14 rods was tested. The mean and standard deviations obtained were 17.85 and 1.955 thousand pounds respectively. Is the result of the experiment significant ? Also obtain the 95 per cent fiducial limits from the sample for the average breaking strength of steel rods.

(b) A sample of 9 shafts is inspected from a production line. The following measurements are the diameters (in mm.) of shafts : 45.010, 45.020, 45.021, 45.015, 45.019, 45.018, 45.020, 45.023 and 45.005. If the production line meets the specifications laid by the I.S.I., with S.D. 0.006 mm, estimate the 95% confidence interval within which the true diameter of the shaft lies.

[Madras Univ. B.E., 1989]

10. a random sample of 8 envelopes is taken from letter box of a post office and their weights in grams are found to be 12.1, 11.9, 12.4, 12.3, 11.9, 12.1, 12.4, 12.1.

(a) Find 99% confidence limits for the mean weight of the envelopes received at that post office.

(b) Using the result of part (a), does this sample indicate at 1% level that the average weight of envelopes received at that post office is 12.35 gms.

11. A random sample of nine from men of a large city gave a mean height 68 inches and the unbiased estimate of the population variance found from the

sample was 4.5 inches. Proceed as far as you can to test for a mean height of 68.5 inches for the men of the city. Also state how you would proceed further.

**14.2.10. t-Test for Difference of Means.** Suppose we want to test if two independent samples  $x_i$  ( $i = 1, 2, \dots, n_1$ ) and  $y_j$  ( $j = 1, 2, \dots, n_2$ ) of sizes  $n_1$  and  $n_2$  have been drawn from two normal populations with means  $\mu_X$  and  $\mu_Y$  respectively.

Under the null hypothesis ( $H_0$ ) that the samples have been drawn from the normal populations with means  $\mu_X$  and  $\mu_Y$  and under the assumption that the population variances are equal, i.e.,  $\sigma_X^2 = \sigma_Y^2 = \sigma^2$  (say), the statistic

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \dots(14.7)$$

where  $\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i, \quad \bar{y} = \frac{1}{n_2} \sum_{j=1}^{n_2} y_j$

and  $S^2 = \frac{1}{n_1 + n_2 - 2} \left[ \sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2 \right] \quad \dots[14.7(a)]$

is an unbiased estimate of the common population variance  $\sigma^2$ , follows Student's *t*-distribution with  $(n_1 + n_2 - 2)$  d.f.

**Proof. Distribution of *t* defined in (14.7).**

$$\xi = \frac{(\bar{x} - \bar{y}) - E(\bar{x} - \bar{y})}{\sqrt{V(\bar{x} - \bar{y})}} \sim N(0, 1)$$

But  $E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_X - \mu_Y$

and  $V(\bar{x} - \bar{y}) = V(\bar{x}) + V(\bar{y})$

[The covariance term vanishes since samples are independent.]

$$= \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \quad (\text{By assumption})$$

$$\therefore \xi = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1) \quad \dots(*)$$

Let  $\chi^2 = \left[ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right] / \sigma^2$

$$= \left[ \sum_i (x_i - \bar{x})^2 / \sigma^2 \right] + \left[ \sum_j (y_j - \bar{y})^2 / \sigma^2 \right] = \frac{n_1 s_X^2}{\sigma^2} + \frac{n_2 s_Y^2}{\sigma^2} \quad \dots(**)$$

Since  $n_1 s_X^2 / \sigma^2$  and  $n_2 s_Y^2 / \sigma^2$  are independent  $\chi^2$ -variates with  $(n_1 - 1)$  and  $(n_2 - 1)$  d.f. respectively, by the additive property of chi-square distribution,  $\chi^2$  defined in (\*\*) is a  $\chi^2$ -variate with  $(n_1 - 1) + (n_2 - 1)$ , i.e.,  $n_1 + n_2 - 2$  d.f.

Further, since sample mean and sample variance are independently distributed,  $\xi$  and  $\chi^2$  are independent random variables.

Hence Fisher's  $t$  statistic is given by

$$\begin{aligned} t &= \frac{\xi}{\sqrt{\frac{\chi^2}{n_1 + n_2 - 2}}} \\ &= \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &\times \frac{1}{\sqrt{\left[ \frac{1}{n_1 + n_2 - 2} \left\{ \sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2 \right\} / \sigma^2 \right]}} \\ &= \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{S \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}, \end{aligned}$$

where  $S^2 = \frac{1}{n_1 + n_2 - 2} \left[ \sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2 \right]$

and it follows Student's  $t$ -distribution with  $(n_1 + n_2 - 2)$  d.f. (c.f. Remark § 14-2-3, page 14-4).

**Remarks 1.**  $S^2$ , defined in 14-7(a) is an unbiased estimate of the common population variance  $\sigma^2$ , since

$$\begin{aligned} E(S^2) &= \frac{1}{n_1 + n_2 - 2} E \left[ \sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2 \right] \\ &= \frac{1}{n_1 + n_2 - 2} E[(n_1 - 1) S_X^2 + (n_2 - 1) S_Y^2] \\ &= \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) E(S_X^2) + (n_2 - 1) E(S_Y^2)] \\ &= \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) \sigma^2 + (n_2 - 1) \sigma^2] = \sigma^2 \end{aligned}$$

**2.** An important deduction which is of much practical utility is discussed below :

Suppose we want to test if : (a) two independent samples  $x_i$  ( $i = 1, 2, \dots, n_1$ ), and  $y_j$  ( $j = 1, 2, \dots, n_2$ ), have been drawn from the populations with same means or (b) the two sample means  $\bar{x}$  and  $\bar{y}$  differ significantly or not.

Under the null hypothesis  $H_0$  that (a) samples have been drawn from the populations with the same means, i.e.,  $\mu_X = \mu_Y$  or (b) the sample means  $\bar{x}$  and  $\bar{y}$  do not differ significantly, [From (14-7) the statistic :

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad [\because \mu_x = \mu_y, \text{ under } H_0] \quad \dots(14.8)$$

where symbols are defined in (14.7a), follows Student's *t*-distribution with  $(n_1 + n_2 - 2)$  d.f.

3. *On the assumption of t-test for difference of means.* Here we make the following three fundamental assumptions :

(i) Parent populations, from which the samples have been drawn are normally distributed.

(ii) The population variances are equal and unknown, i.e.,  $\sigma_x^2 = \sigma_y^2 = \sigma^2$ , (say), where  $\sigma^2$  is unknown.

(iii) The two samples are random and independent of each other.

Thus before applying *t*-test for testing the equality of means it is theoretically desirable to test the equality of population variances by applying *F*-test. If the variances do not come out to be equal then *t*-test becomes invalid and in that case Behren's '*d*'-test based on fiducial intervals is used. For practical problems, however, the assumptions (i) and (ii) are taken for granted.

4. Paired *t*-test For Difference of Means. Let us now consider the case when (i) the sample sizes are equal, i.e.,  $n_1 = n_2 = n$  (say), and (ii) the two samples are not independent but the sample observations are paired together, i.e., the pair of observations  $(x_i, y_i)$ , ( $i = 1, 2, \dots, n$ ) corresponds to the same ( $i$ th) sample unit. The problem is to test if the sample means differ significantly or not.

For example, suppose we want to test the efficacy of a particular drug, say, for inducing sleep. Let  $x_i$  and  $y_i$  ( $i = 1, 2, \dots, n$ ) be the readings, in hours of sleep, on the  $i$ th individual, before and after the drug is given respectively. Here instead of applying the difference of the means test discussed in § 14.2-10, we apply the paired *t*-test given below.

Here we consider the increments,  $d_i = x_i - y_i$ , ( $i = 1, 2, \dots, n$ ).

Under the null hypothesis,  $H_0$  that increments are due to fluctuations of sampling, i.e., the drug is not responsible for these increments, the statistic.

$$t = \frac{\bar{d}}{s/\sqrt{n}} \quad \dots(14.9)$$

where  $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$  and  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$  ...[14.9(a)]

follows Student's *t*-distribution with  $(n - 1)$  d.f.

Example 14.7. Below are given the gain in weights (in lbs.) of pigs fed on two diets A and B.

#### Gain in weight

Diet A : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet B : 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Test, if the two diets differ significantly as regards their effect on increase in weight.

**Solution.** Null hypothesis,  $H_0: \mu_X = \mu_Y$ , i.e., there is no significant difference between the mean increase in weight due to diets A and B.

Alternative hypothesis,  $H_1: \mu_X \neq \mu_Y$  (two-tailed).

Diet A				Diet B			
X	$X - \bar{X}$	$(X - \bar{X})^2$		Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$	
25	-3	9		44	14	196	
32	4	16		34	4	16	
30	2	4		22	-8	64	
34	6	36		10	-20	400	
24	-4	16		47	17	289	
14	-14	196		31	1	1	
32	4	16		40	10	100	
24	-4	16		30	0	0	
30	2	4		32	2	4	
31	3	9		35	5	25	
35	7	49		18	-12	144	
25	-3	9		21	-9	81	
Total	336	0	380	Total	450	0	1410

Under null hypothesis ( $H_0$ ):

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1 + n_2 - 2}$$

Here  $\begin{cases} n_1 = 12, \\ \sum x = 336 \end{cases}$  and  $\begin{cases} n_2 = 15 \\ \sum y = 450 \end{cases}$   
 $\sum (x - \bar{x})^2 = 380$        $\sum (y - \bar{y})^2 = 1410$

$$\therefore \bar{x} = \frac{336}{12} = 28, \bar{y} = \frac{450}{15} = 30$$

and  $S^2 = \frac{1}{n_1 + n_2 - 2} [\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2] = 71.6$

$$\therefore t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{28 - 30}{\sqrt{71.6 \left( \frac{1}{12} + \frac{1}{15} \right)}}$$

$$= \frac{-2}{\sqrt{10.74}} = -0.609$$

Tabulate  $t_{0.05}$  for  $(12 + 15 - 2) = 25$  d.f. is 2.06.

**Conclusion.** Since calculated  $|t|$  is less than tabulated  $t$ ,  $H_0$  may be accepted at 5% level of significance and we may conclude that the two diets do not differ significantly as regards their effect on increase in weight.

**Remark.** Here  $\bar{x}$  and  $\bar{y}$  come out to be integral values and hence the direct method of computing  $\sum(x - \bar{x})^2$  and  $\sum(y - \bar{y})^2$  is used. In case  $\bar{x}$  and (or)  $\bar{y}$  comes out to be fractional, then the step deviation method is recommended for computation of  $\sum(x - \bar{x})^2$  and  $\sum(y - \bar{y})^2$ .

**Example 14-8.** Samples of two types of electric light bulbs were tested for length of life and following data were obtained :

	Type I	Type II
Sample No.	$n_1 = 8$	$n_2 = 7$
Sample Means	$\bar{x}_1 = 1,234$ hrs.	$\bar{x}_2 = 1,036$ hrs.
Sample S.D.'s	$s_1 = 36$ hrs.	$s_2 = 40$ hrs.

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life ?

**Solution.** Null Hypothesis,  $H_0 : \mu_x = \mu_y$ , i.e., the two types I and II of electric bulbs are identical.

Alternative Hypothesis,  $H_1 : \mu_x > \mu_y$ , i.e., type I is superior to type II.

**Test Statistic.** Under  $H_0$ , the test statistic is :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1 + n_2 - 2} = t_{13},$$

where  $S^2 = \frac{1}{n_1 + n_2 - 2} \left[ \sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2 \right]$

$$= \frac{1}{n_1 + n_2 - 2} [n_1 s_1^2 + n_2 s_2^2] = \frac{1}{13} [8 \times (36)^2 + 7 \times (40)^2] = 1659.08$$

$$\therefore t = \frac{1234 - 1036}{\sqrt{1659.08 \left( \frac{1}{8} + \frac{1}{7} \right)}} = \frac{198}{\sqrt{1659.08 \times 0.2679}} = 9.39$$

Tabulated value of  $t$  for 13 d.f. at 5% level of significance for right (single) tailed test is 1.77. [This is the value of  $t_{0.10}$  for 13 d.f. from two-tail tables given in Appendix].

**Conclusion.** Since calculated ' $t$ ' is much greater than tabulated ' $t$ ', it is highly significant and  $H_0$  is rejected. Hence the two types of electric bulbs differ significantly. Further since  $\bar{x}_1$  is much greater than  $\bar{x}_2$ , we conclude that type I is definitely superior to type II.

**Example 14.9.** The heights of six randomly chosen sailors are in inches : 63, 65, 68, 69, 71, and 72. Those of 10 randomly chosen soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Discuss, the light that these data throw on the suggestion that sailors are on the average taller than soldiers.

**Solution.** If the heights of sailors and soldiers be represented by the variables  $X$  and  $Y$  respectively then the Null Hypothesis is,  $H_0 : \mu_X = \mu_Y$ , i.e., the sailors are not on the average taller than the soldiers.

Alternative Hypothesis,  $H_1 : \mu_X > \mu_Y$  (Right-tailed).

Under  $H_0$ , the test statistic is :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1 + n_2 - 2} = t_{14}$$

Sailors

Soldiers

$X$	$d = X - A$ $\doteq X - 68$	$d^2$	$Y$	$D = Y - B$ $\doteq Y - 66$	$D^2$
63	-5	25	61	-5	25
65	-3	9	62	-4	16
68	0	0	65	-1	1
69	1	1	66	0	0
71	3	9	69	3	9
72	4	16	69	3	9
			70	4	16
			71	5	25
Total	0	60	72	6	36
			73	7	49
			Total	18	186

$$\therefore \bar{x} = A + \frac{\sum d}{n_1} \\ = 68 + 0 = 68$$

$$\text{and } \sum(x - \bar{x})^2 = \sum d^2 - \frac{(\sum d)^2}{n_1} \\ = 60 - 0 = 60$$

$$\bar{y} = B + \frac{\sum D}{n_2} \\ = 66 + \frac{18}{10} = 67.8$$

$$\text{and } \sum(y - \bar{y})^2 = \sum D^2 - \frac{(\sum D)^2}{n_2} \\ = 186 - \frac{324}{10} = 153.6$$

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left[ \sum(x - \bar{x})^2 + \sum(y - \bar{y})^2 \right] = \frac{1}{14} (60 + 153.6) = 15.2571$$

$$\therefore t = \frac{68 - 67.8}{\sqrt{15.2571} \left( \frac{1}{6} + \frac{1}{10} \right)^{1/2}} = \frac{0.2}{\sqrt{15.2571} \times 0.2667} = 0.099$$

Tabulated  $t_{0.05}$  for 14 d.f. for single-tail test is 1.76.

*Conclusion.* Since calculated  $t$  is much less than 1.76, it is not at all significant at 5% levels of significance. Hence null hypothesis may be retained at 5% level of significance and we conclude that the data are inconsistent with the suggestion that the sailors are on the average taller than soldiers.

**Example 14-10.** A certain stimulus administered to each of the 12 patients resulted in the following increase of blood pressure :

$$5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4 \text{ and } 6$$

Can it be concluded that the stimulus will, in general, be accompanied by an increase in blood pressure ? [Delhi Univ. B.Sc. 1989]

**Solution.** Here we are given the increments in blood pressure i.e.,

$$d_i (= x_i - y_i).$$

*Null Hypothesis,*  $H_0 : \mu_X = \mu_Y$ , i.e., there is no significant difference in the blood pressure readings of the patients before and after the drug. In other words, the given increments are just by chance (fluctuations of sampling) and not due to the stimulus.

*Alternative Hypothesis,*  $H_1 : \mu_X < \mu_Y$ , i.e., the stimulus results in an increase in blood pressure.

*Test Statistic.* Under  $H_0$ , the test statistic is :

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$$

$d$	5	2	8	-1	3	0	-2	1	5	0	4	6	31
$d^2$	25	4	64	1	9	0	4	1	25	0	16	36	185

$$S^2 = \frac{1}{n-1} \sum (d - \bar{d})^2 = \frac{1}{n-1} \left[ \sum d^2 - \frac{(\sum d)^2}{n} \right]$$

$$= \frac{1}{11} \left[ 185 - \frac{(31)^2}{12} \right] = \frac{1}{11} (185 - 80.08) = 9.5382$$

$$\text{and } \bar{d} = \frac{\sum d}{n} = \frac{31}{12} = 2.58$$

$$\therefore t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2.58 \times \sqrt{12}}{\sqrt{9.5382}} = \frac{2.58 \times 3.464}{3.09} = 2.89$$

Tabulated  $t_{0.05}$  for 11 d.f. for right-tail test is 1.80. [This is the value of  $t_{0.10}$  for 11 d.f. in the Table for two-tailed test given in the Appendix].

**Conclusion.** Since calculated  $t > t_{0.05}$ ,  $H_0$  is rejected at 5% level of significance. Hence we conclude that the stimulus will, in general, be accompanied by an increase in blood pressure.

**Example 14-11.** In a certain experiment to compare two types of pig foods *A* and *B*, the following results of increase in weights were observed in pigs :

Pig number		1	2	3	4	5	6	7	8	Total
Increase in weight in lb	Food A	49	53	51	52	47	50	52	53	407
	Food B	52	55	52	53	50	54	54	53	423

(i) Assuming that the two samples of pigs are independent, can we conclude that food *B* is better than food *A* ?

(ii) Also examine the case when the same set of eight pigs were used in both the foods.

**Solution.** Null Hypothesis,  $H_0$ . If the increase in weights due to foods *A* and *B* are denoted by *X* and *Y* respectively then  $H_0 : \mu_X = \mu_Y$ , i.e., there is no significant difference in increase in weights due to diets *A* and *B*.

Alternative Hypothesis,  $H_1 : \mu_X < \mu_Y$  (Left-tailed).

(i) If the two samples of pigs be assumed to be independent, then we will apply *t*-test for difference of means to test  $H_0$ .

**Test Statistic.** Under  $H_0 : \mu_X = \mu_Y$ , the test criterion is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1 + n_2 - 2}$$

Food A			Food B		
X	d = X - 50	$d^2$	Y	D = Y - 52	$D^2$
49	-1	1	52	0	0
53	3	9	55	3	9
51	1	1	52	0	0
52	2	4	53	1	1
47	-3	9	50	-2	4
50	0	0	54	2	4
52	2	4	54	2	4
53	3	9	53	1	1
	7	37		7	23

$$\therefore \bar{x} = 50 + \frac{7}{8} = 50.875$$

$$\bar{y} = 52 + \frac{7}{8} = 52.875$$

$$\text{and } \left. \begin{aligned} \sum(x - \bar{x})^2 &= \sum d^2 - \frac{(\sum d)^2}{n_1} \\ &= 37 - \frac{49}{8} \\ &= 30.875 \end{aligned} \right\} \quad \left. \begin{aligned} \sum(y - \bar{y})^2 &= \sum D^2 - \frac{(\sum D)^2}{n_2} \\ &= 23 - \frac{49}{8} \\ &= 16.875 \end{aligned} \right\}$$

$$\begin{aligned} S^2 &= \frac{1}{n_1 + n_2 - 2} [\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2] \\ &= \frac{1}{14} (30.875 + 16.875) = 3.41 \end{aligned}$$

$$\therefore t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{50.875 - 52.875}{\sqrt{3.41 \left( \frac{1}{8} + \frac{1}{8} \right)}} = -2.17$$

Tabulated  $t_{0.05}$  for  $(8 + 8 - 2) = 14$  d.f. for one-tail test is 1.76.

**Conclusion.** The critical region for the left-tail test is  $t < -1.76$ . Since calculated  $t$  is less than  $-1.76$ ,  $H_0$  is rejected at 5% level of significance. Hence we conclude that the foods  $A$  and  $B$  differ significantly as regards their effect on increase in weight. Further, since  $\bar{y} > \bar{x}$ , food  $B$  is superior to food  $A$ .

(ii) If the same set of pigs is used in both the cases, then the readings  $X$  and  $Y$  are not independent but they are paired together and we apply the paired  $t$ -test for testing  $H_0$ .

Under  $H_0 : \mu_X = \mu_Y$ , the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$$

$X$	49	53	51	52	47	50	52	53	Total
$Y$	52	55	52	53	50	54	54	53	
$d = X - Y$	-3	-2	-1	-1	-3	-4	-2	0	-16
$d^2$	9	4	1	1	9	16	4	0	44

$$\therefore \bar{d} = \frac{\sum d}{n} = \frac{-16}{8} = -2$$

$$\text{and } S^2 = \frac{1}{n-1} \left[ \sum d^2 - \frac{(\sum d)^2}{n} \right] = \frac{1}{7} \left[ 44 - \frac{256}{8} \right] = 1.714$$

$$\therefore |t| = \frac{|\bar{d}|}{\sqrt{S^2/n}} = \frac{2}{\sqrt{1.7143/8}} = \frac{2}{0.4629} = 4.32$$

Tabulated  $t_{0.05}$  for  $(8 - 1) = 7$  d.f. for one-tail test is 1.90.

*Conclusion.* Here also the observed value of '*t*' is significant at 5% level of significance and we conclude that food *B* is superior to food *A*.

### EXERCISE 14 (c)

1. Explain, stating clearly the assumptions involved, the *t*-test for testing the significance of the difference between the two sample means.

2. Two independent samples of 8 and 7 items respectively had the following values

Sample I...	9	11	13	11	15	9	12	14
Sample II...	10	12	10	14	9	8	10	

Is the difference between the means of samples significant?

3. (a) Two horses *A* and *B* were tested according to the time (in seconds) to run a particular track with the following results :

Horse A : 28	30	32	33	33	29	34
Horse B : 29	30	30	24	27	29	

Test whether the two horses have the same running capacity. [5 per cent values of *t* for 11 and 12 degrees of freedom respectively are 2.20 and 2.18].

Ans. Calculated *t* = 2.5 (approx.)

(b) The gain in weight of two random samples of rats fed on two different diets *A* and *B* are given below. Examine whether the difference in mean increases in weight is significant.

Diet A : 13.	14	10	11	12	16	10	8
Diet B : 7	10	12	8	10	11	9	10

4. (a) Show how you would use Student's *t*-test to decide whether the two sets of observations

[17, 27, 18, 25, 27, 29, 27, 23, 17] and [16, 16, 20, 16, 20, 17, 15, 21] indicate samples drawn from the same universe.

(b) A reading test is given to an elementary school class that consists of 12 Anglo-American children and 10 Mexican-American children. The results of the test are :

Anglo-American	Mexican-American
$\bar{x}_1 = 74$	$\bar{x}_2 = 70$
$s_1 = 8$	$s_2 = 10$

Is the difference between the means of the two groups significant at the 0.05 level? Given  $t_{20} = 2.086$ ,  $t_{22} = 2.074$  at 5% level.

[Delhi Univ. M.C.A., 1986]

5. (a) For a random sample of 10 pigs, fed on a diet *A*, the increases in weight in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 lbs.

For another random sample of 12 pigs fed on diet *B*, the increases in the same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 lbs.

Find if the two samples are significantly different regarding the effect of diet, given that for d.f.  $v = 20, 21, 22$ , the five per cent values of *t* are respectively 2.09, 2.07, 2.06.

Ans. *t* = 1.51; Sample means do not differ significantly.

(b) Two independent samples of rats chosen among both the series had the following increase in weights when fed on a diet. Can you say that the mean increase in weight differs significantly with sex?

Male : 96, 88, 97, 89, 92, 95 and 90

Female : 112, 80, 98, 100, 84, 82, 89, 95, 100 and 96.

6. (a) Ten soldiers visit a rifle range for two consecutive weeks. For the first week their scores are

67, 24, 57, 55, 63, 54, 56, 68, 33, 43

and during the second week they score in the same order—

70, 38, 58, 58, 56, 67, 68, 72, 42, 38

Examine if there is any significant difference in their performance.

(b) Two independent groups of 10 children were tested to find how many digits they could repeat from memory after hearing them. The results are as follows :

Group A : 8 6 5 7 6 8 7 4 5 6

Group B : 10 6 7 8 6 9 7 6 7 7

Is the difference between the mean scores of the two groups significant?

(c) Measurements of the fat content of two kinds of ice cream, Brand A and Brand B, yielded the following sample data :

Brand A : 13.5 14.0 13.6 12.9 13.0

Brand B : 12.9 13.0 12.4 13.5 12.7

Test the null hypothesis  $\mu_1 = \mu_2$ , (where  $\mu_1$  and  $\mu_2$  are the respective true average fat contents of the two kinds of ice cream), against the alternative hypothesis  $\mu_1 \neq \mu_2$  at the level of significance  $\alpha = 0.05$ .

[Madras Univ. B.E., 1990]

7. (a) A random sample of 16 values from a normal population has a mean of 41.5 inches and sum of squares of deviations from the mean is equal to 135 inches. Another sample of 20 values from an unknown population has a mean of 43.0 inches and sum of squares of deviations from their mean is equal to 171 inches. Show that the two samples may be regarded as coming from the same normal population.

(b) A company is interested in knowing if there is a difference in the average salary received by foremen in two divisions. Accordingly samples of 12 foremen in the first division and 10 foremen in the second division are selected at random. Based upon experience, foremen's salaries are known to be approximately normally distributed, and the standard deviations are about the same.

	First Division	Second division
Sample size	12	10
Average monthly salary of foremen (Rs.)	1,050	980
Standard deviation of salaries (Rs.)	68	74

The table value of  $t$  for 20 d.f. at 5% level of significance is 2.086.

Ans.  $t = 2.2$ . Reject  $H_0 : \mu_x = \mu_y$

(c) The average number of articles produced by two machines per day are 200 and 250 with standard deviations 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 1% level of significance ?

**Ans.**  $t = -7.65$ . **Hint.** Here  $n_1 = n_2 = 25$ .

8. Eleven school boys were given a test in Statistics. They were given a month's tuition and a second test was held at the end of it. Do the marks give evidence that the students have benefited by the extra coaching ?

Boys	1	2	3	4	5	6	7	8	9	10	11
Marks in 1st test	23	20	19	21	18	20	18	17	23	16	19
Marks in 2nd test	24	19	22	18	20	22	20	20	23	20	18

**Ans.**  $H_0: \mu_1 = \mu_2$ ;  $H_1: \mu_1 < \mu_2$ . Paired  $|t| = 1.483$ . Not significant. Hence, students have not benefited from extra coaching.

9.(a) The following table gives the additional hours of sleep gained by 10 patients in an experiment to test the effect of a drug. Do these data give evidence that the drug produces additional hours of sleep ?

Patients	1	2	3	4	5	6	7	8	9	10
Hours gained :	0.7	0.1	0.2	1.2	0.31	0.4	3.7	0.8	3.8	2.0

(b) A drug was administered to 10 patients, and the increments in their blood pressure were recorded to be 6, 3, -2, 4, -3, 4, 6, 0, 3, 2. Is it reasonable to believe that the drug has no effect on change of blood pressure ? Use 5% significance level, and assume that for 9 degrees of freedom,  $P(t > 2.26) = 0.025$ . [Calcutta Univ. B.Sc.(Maths. Hons.), 1986]

(c) The scores of 10 candidates prior and after training are given below :

Prior :	84	48	36	37	54	69	83	96	90	65
After :	90	58	56	49	62	81	84	86	84	75

Is the training effective ? [Calicut Univ. B.Sc., Oct. 1992]

10. The following table gives measurements of blood pressure on subjects by two investigators :

Subject No. :	1	2	3	4	5	6	7	8	9	10
Investigator I :	70	68	56	75	80	90	68	75	56	58
Investigator II :	68	70	52	73	75	78	67	70	54	55

No other details of the experiment were given.

(i) If a valid inference has to be drawn about the difference between the investigators, mention the precautions that should have been taken in conducting the experiment with respect to the time of measurement, interval between the first and second measurements, the order in which the investigators measure, etc.

(ii) After the experiment was conducted it was discovered that all the subjects were unrelated except that No. 10 was the father of No. 9. Assuming that all the precautions you mention in (a) are satisfied, analyse the data to draw an inference on the difference between the investigators. 5 per cent values of the *t*-statistic corresponding to various degrees of freedom are as follows :

5 per cent values of <i>t</i> ...	2.40	2.31	2.26	2.23	2.10	2.09
Degrees of freedom...	7	8	9	10	18	19

11. The following are the values of the cephalic index found in two samples of skulls, one consisting of 15 and the other of 13 individuals.

Sample I : 74.1 77.7 74.0 74.4 73.8 79.3 75.8 82.8  
72.2 75.2 78.2 77.1 78.4 76.3 76.8

Sample II : 70.8 74.9 74.2 70.4 69.2 72.2 76.8 72.4  
77.4 78.1 72.8 74.3 74.7

(i) Test the hypothesis that the means of population I and population II could be equal.

(ii) Is it possible that the sample II has come from a population of mean 72.0?

(iii) Obtain confidence limits for the mean of population I and for the mean of population II.

(Assume that the distribution of cephalic indices for a homogeneous population is normal.)

12. (a) The following table gives the gain in weight in decagrams in a feeding experiment with pigs on the relative value of limestone and bone meal for bone development.

Limestone	49.2	53.3	50.6	52.0	46.8	50.5	52.1	53.0
Bone meal	51.5	54.9	52.2	53.3	51.6	54.1	54.2	53.3

Test for the significance of difference between the means in two ways :

(i) by assuming that the values are paired.

(ii) by assuming that the values are not paired.

(b) The following table shows the mean number of bacterial colonies per plate obtainable by four slightly different methods from soil samples taken at 4 P.M. and 8 P.M. respectively.

Method	A	B	C	D
4 P.M.	29.75	27.50	30.25	27.80
8 P.M.	39.20	40.60	36.20	42.40

Are there significantly more bacteria at 8 P.M. than at 4 P.M.?

[Given  $t_{0.05}(3) = 3.18$  and  $t_{0.01}(3) = 5.84$ ]

13. (a) It is believed that glucose treatment will extend the sleep time of mice. In an experiment to test this hypothesis ten mice selected at random are given glucose treatment and are found to have a mean hexobarbital sleep time of 47.2 min with a standard deviation of 9.3 min. A further sample of ten untreated mice are found to have a mean hexobarbital sleep time of 28.5 min. with a standard deviation of 7.2 min. Are these results significant evidence in favour of the hypothesis?

Find 95% confidence limits for the population mean difference in sleep time. State any assumptions made concerning the data in carrying out the test and finding the limits. [Bangalore Univ. B.E., Oct. 1992]

(b) An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material I were tested, by exposing each piece to a machine measuring wear. Ten pieces of material II were similarly tested. In each case the depth of wear was observed. The sample of material I gave an average (coded) wear 8.5 units with a standard deviation of 0.4

while the sample of material II gave an average of 8.1 and a standard deviation of 0.5. Test the hypothesis that the two types of material exhibit the same mean abrasive wear at the 0.10 level of significance. Assume the populations to be approximately normal with equal variances.

If the level of significance is 0.01, what will be your conclusion?

[*Delhi Univ. M.E., 1992*]

**14.2.11. *t*-test For Testing Significance of an Observed sample Correlation Coefficient.** If  $r$  is the observed correlation coefficient in a sample of  $n$  pairs of observations from a bivariate normal population, then Prof. Fisher proved that under the null hypothesis  $H_0 : \rho = 0$ , i.e., *population correlation coefficient is zero*, the statistic :

$$t = \frac{r}{\sqrt{(1 - r^2)}} \sqrt{(n - 2)} \quad \dots(14.9)$$

follows Student's *t*-distribution with  $(n - 2)$  d.f. (c.f. Remark to § 14.3 page 14.41).

If the value of  $t$  comes out to be significant, we reject  $H_0$  at the level of significance adopted and conclude that  $\rho \neq 0$ , i.e., ' $r$ ' is significant of correlation in the population.

If  $t$  comes out to be non-significant then  $H_0$  may be accepted and we conclude that variables may be regarded as uncorrelated in the population.

**Example 14.12.** A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of 0.6. Is this significant of correlation in the population?

**Solution.** We set up the null hypothesis,  $H_0 : \rho = 0$ , i.e., the observed sample correlation coefficient is not significant of any correlation in the population.

$$\text{Under } H_0 : t = \frac{r \sqrt{(n - 2)}}{\sqrt{(1 - r^2)}} \sim t_{(n-2)}$$

$$\text{Here } t = \frac{0.6 \sqrt{27 - 2}}{\sqrt{(1 - 0.36)}} = \frac{3}{\sqrt{0.64}} = 3.75$$

Tabulated  $t_{0.05}$  for  $(27 - 2) = 25$  d.f. is 2.06.

**Conclusion.** Since calculated  $t$  is much greater than the tabulated  $t$ , it is significant and hence  $H_0$  is discredited at 5% level of significance. Thus we conclude that the variables are correlated in the population.

**Example 14.13.** Find the least value of  $r$  in a sample of 18 pairs of observations from a bi-variate normal population, significant at 5% level of significance.

**Solution.** Here  $n = 18$ . From the tables  $t_{0.05}$  for  $(18 - 2) = 16$  d.f. is 2.12

$$\text{Under } H_0 : \rho = 0, \quad t = \frac{r \sqrt{(n - 2)}}{\sqrt{(1 - r^2)}} \sim t_{(n-2)}$$

In order that the calculated value of  $t$  is significant at 5% level of significance, we should have

$$\begin{aligned} \left| \frac{r \sqrt{(n-2)}}{\sqrt{1-r^2}} \right| > t_{0.05} &\Rightarrow \left| \frac{r \sqrt{16}}{\sqrt{1-r^2}} \right| > 2.12 \\ \Rightarrow 16r^2 > (2.12)^2(1-r^2) &\Rightarrow 20.493r^2 > 4.493 \\ \Rightarrow r^2 > \frac{4.493}{20.493} &= 0.2192 \end{aligned}$$

Hence  $|r| > 0.4682$

**Example 14.14.** A coefficient of correlation of 0.2 is derived from a random sample of 625 pairs of observations. (i) Is this value of  $r$  significant? (ii) What are the 95% and 99% confidence limits to the correlation coefficient in the population?

**Solution.** Under the null hypothesis  $H_0 : \rho = 0$ , i.e., the value of  $r = 0.2$  is not significant; the test statistics is :

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

$$\text{Now } t = \frac{0.2 \times \sqrt{(625-2)}}{\sqrt{1-0.04}} = 5.09$$

Since  $d.f. = 625 - 2 = 623$ , the significant values of  $t$  are same as in the case of normal distribution, viz.,  $t_{0.05} = 1.96$  and  $t_{0.01} = 2.58$ . Since calculated  $t$  is much greater than these values; it is highly significant. Hence  $H_0 : \rho = 0$  is rejected and we conclude that the sample correlation is significant of correlation in the population.

95% Confidence Limits for  $\rho$  (population correlation coefficient) are

$$\begin{aligned} r \pm 1.96 \text{ S.E.}(r) &= r \pm 1.96 (1-r^2)/\sqrt{n} \quad [\text{Since } n \text{ large}] \\ &= 0.2 \pm (1.96 \times 0.96/\sqrt{625}) \\ &= 0.2 \pm 0.075 = (0.125, 0.275) \end{aligned}$$

99% Confidence Limits for  $\rho$  are :

$$0.2 \pm 2.58 \times 0.0384 = 0.2 \pm 0.099 = (0.101, 0.299)$$

### EXERCISE 14 (d)

1. A restaurant owner ranked his 17 waiters in terms of their speed and efficiency on the job. He correlated these ranks with the total amount of tips each of these waiters received for a one-week period. The obtained value of correlation coefficient is 0.438. What do you conclude?

Giyen :  $t_{15}(0.05) = 2.131$ ,  $t_{16}(0.05) = 2.120$  for two-tailed test.

[Delhi Univ. M.C.A., 1990]

2. Test the significance of the values of correlation coefficient ' $r$ ' obtained from samples of size  $n$  pairs from a bivariate normal population.

(i)  $r = 0.6$ ,  $n = 38$       (ii)  $r = 0.5$ ,  $n = 11$

**Ans.** (i)  $t = 4.5$ ; Significant at 5% level;  $H_0 : \rho = 0$  rejected.

(ii)  $t = 1.73$ ; Not significant at 5% level.

- (i) Consistent Statistic
- (ii) Unbiased Statistic
- (iii) Sufficient Statistic
- (iv) Efficiency.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1982*]

2. What do you understand by Point Estimation ? When would you say that estimate of a parameter is good ? In particular, discuss the requirements of consistency and unbiasedness of an estimate. Give an example to show that a consistent estimate need not be unbiased.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986*]

3. Discuss the terms (i) estimate, (ii) consistent estimate, (iii) unbiased estimate, of a parameter and show that sample mean is both consistent and unbiased estimate of the population mean.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1986*]

4. (a) If  $s_1^2, s_2^2, \dots, s_r^2$  are  $r$  sample variances based on random samples of sizes  $n_1, n_2, \dots, n_r$ , respectively, and if  $T$  is some statistic given by

$$T = \frac{n_1 s_1^2 + n_2 s_2^2 + \dots + n_r s_r^2}{a},$$

for estimating  $\sigma^2$  as an unbiased estimator, find the value of  $a$ , supposing population is very large and for every sample

$$s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

**Ans.**  $a = (n_1 + n_2 + \dots + n_r) - r$ .

(b) If  $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots, \bar{X}_r$  are the sample means based on samples of sizes  $n_1, n_2, n_3, \dots, n_r$ , respectively, an unbiased estimator,

$$t = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \dots + n_r \bar{X}_r}{k}$$

has been defined to estimate  $\mu$ . Find the value of  $k$ .

**Ans.**  $k = n_1 + n_2 + \dots + n_r$ .

5. (a) For the geometric distribution,

$$f(x, \theta) = \theta (1 - \theta)^{x-1}, (x = 1, 2, \dots), 0 < \theta < 1,$$

Obtain an unbiased estimator of  $1/\theta$ .

[Ans.  $E(\bar{X}) = 1/\theta$ .]

(b) The random variable  $X$  takes the values 1 and 0 with respective probabilities  $\theta$  and  $1 - \theta$ . Independent observations  $X_1, X_2, \dots, X_n$  on  $X$  are available. Write  $\xi = X_1 + X_2 + \dots + X_n$ .

Show that  $\xi(n - \xi)/n(n - 1)$  is an unbiased estimate of  $\theta(1 - \theta)$ .

6. Show that if  $T$  is an unbiased estimator of a parameter  $\theta$ , then  $\lambda_1 T + \lambda_2$  is an unbiased estimator of  $\lambda_1 \theta + \lambda_2$ , where  $\lambda_1$  and  $\lambda_2$  are known constants, but  $T^2$  is a biased estimator of  $\theta^2$ .

7. For the following cases determine if the given estimator is unbiased for the parametric function. When it is biased, derive an unbiased estimator from it.  $\bar{x}$  is the sample mean.

**Proof.** Let  $(x_i, y_i)$ ,  $(i = 1, 2, \dots, n)$  be a random sample of size  $n$  drawn from an uncorrelated bivariate normal population ( $\rho = 0$ ) in which  $E(X) = E(Y) = 0$  and  $V(X) = \sigma_X^2$ ,  $V(Y) = \sigma_Y^2$ . Let the variable  $Y$  be transformed to the variable  $Z$  by means of a linear orthogonal transformation, viz.,

$$\mathbf{Z} = \mathbf{CY}$$

where  $\mathbf{Z}_{n \times 1} = (z_1, z_2, \dots, z_n)'$ ,  $\mathbf{Y}_{n \times 1} = (y_1, y_2, \dots, y_n)'$  and  $\mathbf{C}_{n \times n} = (c_{ij})$ ,  $\mathbf{C}$  is an orthogonal matrix. Let us, in particular, take

$$c_{11} = c_{12} = \dots = c_{1n} = 1/\sqrt{n},$$

so that

$$z_1 = \frac{1}{\sqrt{n}}(y_1 + y_2 + \dots + y_n) = \sqrt{n}\bar{y}$$

Now proceeding as in (Theorem 13.5), we get

$$\sum_{i=2}^n z_i^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = ns_Y^2$$

Since in a bivariate normal distribution, the marginal distributions of  $X$  and  $Y$  are also normal, we have  $Y \sim N(0, \sigma_Y^2)$ . Hence by Fisher's Lemma (Theorem 13.4)  $z_i$ ,  $(i = 1, 2, \dots, n)$  are independent  $N(0, \sigma_Y^2)$ .

$$\text{Now } r = \frac{\text{Cov}(X, Y)}{s_X s_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n s_X s_Y}$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^n (x_i - \bar{x})}{n s_X s_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{n s_X s_Y}$$

$$\therefore \sqrt{n} s_Y r = \frac{\sum (x_i - \bar{x}) y_i}{\sqrt{n} s_X} = z_2, \text{ (say)}, \quad \dots (**)$$

[since the sum of the squares of coefficients of  $y_1, y_2, \dots, y_n$  in (\*\*) is unity.]  
From (\*) and (\*\*), we get

$$ns_Y^2 = \sum_{i=2}^n z_i^2 = \sum_{i=3}^n z_i^2 + z_2^2 = \sum_{i=3}^n z_i^2 + nr^2 s_Y^2$$

$$\Rightarrow (1 - r^2) n s_Y^2 = \sum_{i=3}^n z_i^2 \quad \dots (***)$$

Since  $z_i$ ,  $(i = 1, 2, \dots, n)$  are independent  $N(0, \sigma_Y^2)$ ;

$\therefore (z_i/\sigma_Y)$ ,  $(i = 1, 2, \dots, n)$  are independent  $N(0, 1)$ .

Hence from (\*\*),

$$U = \frac{z_2^2}{\sigma_Y^2} = \frac{nr^2 s_Y^2}{\sigma_Y^2}$$

being the square of a standard normal variate is a  $\chi^2$ -variate with 1 d.f. and from (\*\*\*)

$$V = \sum_{i=3}^n z_i^2 / \sigma_Y^2 = \sum_{i=3}^n (z_i / \sigma_Y)^2 = \frac{(1 - r^2) n s_Y^2}{\sigma_Y^2},$$

being the sum of squares of  $(n - 2)$  independent standard normal variates is an independent  $\chi^2$ -variante with  $(n - 2)$  d.f.

Further, since  $z_2$  and  $(z_3, z_4, \dots, z_n)$  are independent r.v.'s,  $U$  and  $V$  are independent chi-square variates with 1 and  $(n - 2)$  d.f. respectively.

$$\therefore \frac{U}{U + V} = \frac{nr^2 s_Y^2 / \sigma_Y^2}{[nr^2 s_Y^2 + (1 - r^2) n s_Y^2] / \sigma_Y^2} \sim \beta_1 \left( \frac{1}{2}, \frac{n-2}{2} \right)$$

[c.f. Theorem 13.2]

$$\Rightarrow r^2 \sim \beta_1 \left( \frac{1}{2}, \frac{n-2}{2} \right)$$

Hence the probability function of  $r^2$  is given by

$$dF(r^2) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} (r^2)^{(1/2)-1} [1 - r^2]^{\frac{n-2}{2}-1} d(r^2), \quad 0 \leq r^2 \leq 1$$

$$\Rightarrow dF(r) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} [1 - r^2]^{(n-4)/2} dr, \quad -1 \leq r \leq 1$$

the factor 2 disappearing from the fact that total probability in the range  $-1 \leq r \leq 1$  must be unity.

**Remark.** If  $\rho = 0$ , then  $t = \frac{r}{\sqrt{1 - r^2}} \sqrt{(n - 2)}$  is distributed as Student's  $t$  with  $(n - 2)$  d.f.

$$\text{Proof.} \quad t = \frac{r \sqrt{(n-2)}}{\sqrt{1-r^2}} \quad \dots (*)$$

$$\Rightarrow \frac{t^2}{n-2} = \frac{r^2}{1-r^2} = \frac{1}{1-r^2} - 1$$

$$\Rightarrow (1-r^2) = \left[ 1 + \frac{t^2}{n-2} \right]^{-1} \quad \dots (**)$$

From (\*),

$$dt = \sqrt{(n-2)} d[r/\sqrt{1-r^2}]$$

$$dt = \sqrt{(n-2)} \left[ \frac{dr}{\sqrt{1-r^2}} + \left( -\frac{r}{2} \right) \frac{(-2r)dr}{(1-r^2)^{3/2}} \right]$$

$$\Rightarrow dt = \sqrt{(n-2)} \frac{dr}{\sqrt{1-r^2}} \left[ 1 + \frac{r^2}{1-r^2} \right]$$

$$\Rightarrow dt = \sqrt{(n-2)} \times \frac{dr}{(1-r^2)^{3/2}}, \text{ i.e., } dr = \frac{1}{\sqrt{(n-2)}} (1-r^2)^{3/2} dt$$

As  $r$  ranges from  $-1$  to  $1$ , from  $(*)$ ,  $t$  ranges from  $-\infty$  to  $\infty$ .

When  $\rho = 0$ , the p.d.f. of ' $r$ ' is given by (14-12) and it transforms to

$$\begin{aligned} dG(t) &= \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} [1-r^2]^{(n-4)/2} \frac{1}{\sqrt{(n-2)}} (1-r^2)^{3/2} dt \\ &= \frac{1}{\sqrt{(n-2)} B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{n-2}\right]^{(n-1)/2}} , \\ &\quad [\text{From } (**)] \\ &= \frac{1}{\sqrt{(n-2)} B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{n-2}\right]^{(n-2+1)/2}} , \\ &\quad -\infty < t < \infty \end{aligned}$$

which is the p.d.f. of  $t$ -distribution with  $(n-2)$ d.f.

$$\text{Hence } t = \frac{r}{\sqrt{1-r^2}} \sim t_{(n-2)}$$

**Example 14-15.** (a) If  $(x_i, y_i)$  is a random sample drawn from an uncorrelated bivariate normal population, derive the distribution of

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

(b) Further, when  $n = 5$  and if  $P(|r| \geq C) = \alpha$ , show that  $C$  is a root of the equation,

$$C\sqrt{1-C^2} + \sin^{-1} C + \frac{\pi(\alpha-1)}{2} = 0$$

**Solution.** (a) c.f. § 14-3.

$$(b) P(|r| \geq C) = 1 - P(|r| \leq C) = 1 - P(-C \leq r \leq C)$$

$$= 1 - 2P(0 \leq r \leq C) = 1 - 2 \int_0^C f(r) dr$$

[ $\because f(r)$  is symmetrical about  $r = 0$ ]

$$\text{When } n = 5, \quad f(r) = \frac{1}{B\left(\frac{1}{2}, \frac{3}{2}\right)} (1-r^2)^{\frac{1}{2}} dr \quad [\text{c.f. Equation (14-12)}]$$

$$\therefore P(|r| \geq C) = 1 - 2 \frac{\Gamma(2)}{\Gamma(1/2)\Gamma(3/2)} \int_0^C (1-r^2)^{1/2} dr$$

$$= 1 - 2 \times \frac{1}{\frac{1}{2}\pi} \left[ \frac{1}{2} r (1-r^2)^{1/2} + \frac{1}{2} \sin^{-1} r \right]_0^C$$

$$= 1 - \frac{4}{\pi} \left[ \frac{1}{2} C (1-C^2)^{1/2} + \frac{1}{2} \sin^{-1} C \right] = \alpha, \quad (\text{Given})$$

$$\therefore 1 - \frac{2}{\pi} \left[ C(1 - C^2)^{\frac{1}{2}} + \sin^{-1} C \right] = \alpha$$

$$\Rightarrow C(1 - C^2)^{1/2} + \sin^{-1} C + (\alpha - 1) \frac{\pi}{2} = 0$$

**14.4. Non-central *t*-distribution.** The non-central *t*-distribution is the distribution of the ratio of a normal variate with possibly non-zero mean and variance unity, to the square root of an independent  $\chi^2$ -variate divided by its degrees of freedom. If  $X \sim N(\mu, 1)$  and  $Y$  is an independent  $\chi^2$ -variate with  $n$  d.f., then

$$t' = \frac{X}{\sqrt{Y/n}} \quad \dots(14.13)$$

is said to have a non-central *t*-distribution with  $n$  d.f. and non-centrality parameter  $\mu$ . Non-central *t*-distribution is required for the power functions of certain tests concerning normal population.

**p.d.f. of '*t'*.** Since  $X \sim N(\mu, 1)$ , its p.d.f.  $f(\cdot)$  is

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2}(x - \mu)^2 \right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2}(\mu^2 + x^2) \right] \sum_{i=0}^{\infty} \frac{(\mu x)^i}{i!}, -\infty < x < \infty$$

Since  $Y \sim \chi^2_{(n)}$ , its p.d.f.  $g(\cdot)$  is

$$g(y) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-y/2} y^{(n/2)-1}, 0 < y < \infty$$

Since  $X$  and  $Y$  are independent, their joint p.d.f. becomes

$$f(x, y) = \frac{1}{\sqrt{2\pi} 2^{n/2} \Gamma(n/2)} \exp \left[ -\frac{1}{2}(\mu^2 + x^2 + y) \right] y^{(n/2)-1} \sum_{i=0}^{\infty} \frac{(\mu x)^i}{i!}$$

Let us transform to new variables  $t'$  and  $z$  by the substitution :

$$t' = \frac{x}{\sqrt{y/n}} = \frac{\sqrt{n}x}{\sqrt{y}}, z = +\sqrt{y}$$

$$\Rightarrow x = zt' \sqrt{n}, y = z^2$$

Jacobian of transformation  $J$  is

$$J = \begin{vmatrix} \frac{\partial x}{\partial t'} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial t'} & \frac{\partial y}{\partial z} \end{vmatrix} = \begin{vmatrix} \frac{z}{\sqrt{n}} & \frac{t'}{\sqrt{n}} \\ 0 & 2z \end{vmatrix} = \frac{2z^2}{\sqrt{n}}$$

The joint p.d.f. of  $t'$  and  $z$  becomes

$$h(t', z) = \frac{\exp(-\mu^2/2)}{\sqrt{2\pi} 2^{n/2} \Gamma(n/2)} (z^2)^{\frac{n}{2}-1} \sum_{i=0}^{\infty} \frac{(\mu z t' / \sqrt{n})^i}{i!} \cdot \frac{2z^2}{\sqrt{n}}$$

$$\times \exp \left[ -\frac{1}{2} \left( 1 + \frac{t'^2}{n} \right) z^2 \right]; \quad -\infty < t' < \infty, \quad 0 < z < \infty$$

$$= \frac{\exp(-\mu^2/2)}{\sqrt{\pi} 2^{(n-1)/2} \Gamma(n/2)} \sum_{i=0}^{\infty} \left[ \frac{(\mu t')^i}{i! n^{(i+1)/2}} \cdot \exp \left\{ -\frac{1}{2} \left( 1 + \frac{t'^2}{n} \right) z^2 \right\} z^{n+i} \right]$$

Integrating w.r.t.  $z$  in the range 0 to  $\infty$ , we get the p.d.f. of  $t'$

$$h_1(t') = \frac{\exp(-\mu^2/2)}{\sqrt{\pi} 2^{(n-1)/2} \Gamma(n/2)} \times \sum_{i=0}^{\infty} \left[ \frac{(\mu t')^i}{i! n^{(i+1)/2}} \int_0^{\infty} \exp \left\{ -\frac{1}{2} \left( 1 + \frac{t'^2}{n} \right) z^2 \right\} z^{n+i} dz \right]$$

$$= \frac{\exp(-\mu^2/2)}{\sqrt{\pi} 2^{(n-1)/2} \Gamma(n/2)} \times \sum_{i=0}^{\infty} \left[ \frac{(\mu t')^i}{i! n^{(i+1)/2}} \int_0^{\infty} \exp \left\{ -\left( 1 + \frac{t'^2}{n} \right) v \right\} (2v)^{(n+i-1)/2} dv \right]$$

$$= \frac{\exp(-\mu^2/2)}{\sqrt{\pi} \Gamma(n/2)} \sum_{i=0}^{\infty} \left[ \frac{\mu^i 2^{i/2} \Gamma\left(\frac{n+i+1}{2}\right)}{i! n^{(i+1)/2}} \frac{t'^i}{\left(1 + \frac{t'^2}{n}\right)^{(n+i+1)/2}} \right]$$

... [14.13(a)]

which is the p.d.f. of non-central  $t$ -distribution with  $n$  d.f. and non-centrality element  $\mu$ .

**Remark.** If  $\mu = 0$ , we get from [14.13 (a)]

$$h_1(t') = \frac{1}{\sqrt{\pi} \Gamma(n/2)} \cdot \frac{\Gamma((n+1)/2)}{\sqrt{n}} \cdot \frac{1}{\left[1 + \frac{t'^2}{n}\right]^{(n+1)/2}}$$

$$= \frac{1}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)} \left[1 + \frac{t'^2}{n}\right]^{-(n+1)/2}, \quad -\infty < t' < \infty$$

which is the p.d.f. of central  $t$ -distribution with  $n$  d.f.

**14.5. F-statistic. Definition.** If  $X$  and  $Y$  are two independent chi-square variates with  $v_1$  and  $v_2$  d.f. respectively, then F-statistic is defined by

$$F = \frac{X/v_1}{Y/v_2} \quad \dots (14.14)$$

In other words,  $F$  is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's  $F$ -distribution with  $(v_1, v_2)$  d.f. with probability function given by

$$f(F) = \frac{\left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}}}{B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \cdot \frac{\frac{v_1}{F^2} - 1}{\left[1 + \frac{v_1}{v_2}F\right]^{(v_1+v_2)/2}}, \quad 0 \leq F < \infty \quad \dots [14.14(a)]$$

**Remarks 1.** The sampling distribution of  $F$ -statistic does not involve any population parameters and depends only on the degrees of freedom  $v_1$  and  $v_2$ .

**2.** A statistic  $F$  following Snedecor's  $F$ -distribution with  $(v_1, v_2)$  d.f. will be denoted by  $F \sim F(v_1, v_2)$ .

**14.5.1 Derivation of Snedecor's  $F$ -distribution.** Since  $X$  and  $Y$  are independent chi-square variates with  $v_1$  and  $v_2$  d.f. respectively, their joint probability differential is given by

$$\begin{aligned} dF(x, y) &= \left\{ \frac{1}{2^{v_1/2} \Gamma(v_1/2)} \exp(-x/2) x^{(v_1/2)-1} dx \right\} \\ &\quad \times \left\{ \frac{1}{2^{v_2/2} \Gamma(v_2/2)} \exp(-y/2) y^{(v_2/2)-1} dy \right\} \\ &= \frac{1}{2^{(v_1+v_2)/2} \Gamma(v_1/2) \Gamma(v_2/2)} \exp\{- (x+y)/2\} \\ &\quad \times x^{(v_1/2)-1} y^{(v_2/2)-1} dx dy, \quad 0 \leq (x, y) < \infty \end{aligned}$$

Let us make the following transformation of variables :

$$F = \frac{x/v_1}{y/v_2} \text{ and } u = y, \text{ so that } 0 \leq F < \infty, 0 < u < \infty$$

$$\therefore x = \frac{v_1}{v_2} F u = \frac{v_1}{v_2} F u \quad \text{and} \quad y = u$$

Jacobian of transformation  $J$  is given by

$$J = \frac{\partial(x, y)}{\partial(F, u)} = \begin{vmatrix} \frac{v_1}{v_2} u & 0 \\ \frac{v_1}{v_2} F & 1 \end{vmatrix} = \frac{v_1 u}{v_2}$$

Thus the distribution of the transformed variable is

$$\begin{aligned} dG(F, u) &= \frac{1}{2^{(v_1+v_2)/2} \Gamma(v_1/2) \Gamma(v_2/2)} \exp\left\{-\frac{u}{2} \left(1 + \frac{v_1}{v_2} F\right)\right\} \\ &\quad \times \left(\frac{v_1}{v_2} F u\right)^{(v_1/2)-1} u^{(v_2/2)-1} |J| du dF \\ &= \frac{(v_1/v_2)^{v_1/2}}{2^{(v_1+v_2)/2} \Gamma(v_1/2) \Gamma(v_2/2)} \exp\left\{-\frac{u}{2} \left(1 + \frac{v_1}{v_2} F\right)\right\} \\ &\quad \times u^{(v_1+v_2)/2 - 1} F^{(v_1/2)-1} du dF; \quad 0 < u < \infty, 0 \leq F < \infty \end{aligned}$$

Integrating out  $u$  over the range 0 to  $\infty$ , the distribution of  $F$  becomes

$$\begin{aligned} g_1(F) dF &= \frac{(\nu_1/\nu_2)^{(\nu_1/2)} F^{(\nu_1/2)-1}}{2^{(\nu_1+\nu_2)/2} \Gamma(\nu_1/2) \Gamma(\nu_2/2)} \\ &\quad \times \left[ \int_0^\infty \exp \left\{ -\frac{u}{2} \left( 1 + \frac{\nu_1}{\nu_2} F \right) \right\} u^{((\nu_1+\nu_2)/2)-1} du \right] \\ &= \frac{(\nu_1/\nu_2)^{(\nu_1/2)} F^{(\nu_1/2)-1}}{2^{(\nu_1+\nu_2)/2} \Gamma(\nu_1/2) \Gamma(\nu_2/2)} \times \frac{\Gamma[(\nu_1+\nu_2)/2]}{\left[ \frac{1}{2} \left( 1 + \frac{\nu_1}{\nu_2} F \right) \right]^{(\nu_1+\nu_2)/2}} dF \\ \therefore g_1(F) &= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left[ 1 + \frac{\nu_1}{\nu_2} F \right]^{(\nu_1+\nu_2)/2}}, \quad 0 \leq F < \infty \end{aligned}$$

which is the required probability function of  $F$ -distribution with  $(\nu_1, \nu_2)$  d.f.

**Aliter**  $F = \frac{x/\nu_1}{y/\nu_2}$

$\therefore \frac{\nu_1}{\nu_2} F = \frac{x}{y}$ , being the ratio of two independent chi-square variates with  $\nu_1$  and  $\nu_2$  d.f. respectively is a  $\beta_2\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)$  variate. Hence the probability function of  $F$  is given by

$$\begin{aligned} dP(F) &= \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{\left(\frac{\nu_1}{\nu_2} F\right)^{(\nu_1/2)-1}}{\left[ 1 + \frac{\nu_1}{\nu_2} F \right]^{(\nu_1+\nu_2)/2}} d\left(\frac{\nu_1}{\nu_2} F\right) \\ &= \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left[ 1 + \frac{\nu_1}{\nu_2} F \right]^{(\nu_1+\nu_2)/2}} dF, \quad 0 \leq F < \infty \end{aligned}$$

#### 14-5-2. Constants of F-distribution.

$$\begin{aligned} \mu'_{r'} (\text{about origin}) &= E(F^r) = \int_0^\infty F^r f(F) dF \\ &= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty F^r \frac{F^{(\nu_1/2)-1}}{\left[ 1 + \frac{\nu_1}{\nu_2} F \right]^{(\nu_1+\nu_2)/2}} dF \quad ...(*) \end{aligned}$$

To evaluate the integral, put

$$\frac{\nu_1}{\nu_2} F = y, \text{ so that } dF = \frac{\nu_2}{\nu_1} dy$$

$$\begin{aligned} \therefore \mu_r' &= \frac{[\nu_1/\nu_2]^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{\left(\frac{\nu_2}{\nu_1} y\right)^r + (\nu_1/2) - 1}{[1+y]^{(\nu_1+\nu_2)/2}} \left(\frac{\nu_2}{\nu_1}\right) dy \\ &= \frac{\left(\frac{\nu_2}{\nu_1}\right)^r}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{y^r + (\nu_1/2) - 1}{[1+y]^{(\nu_1/2)+r+(\nu_2/2)-r}} dy \\ &= \left(\frac{\nu_2}{\nu_1}\right)^r \cdot \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot B\left(r + \frac{\nu_1}{2}, \frac{\nu_2}{2} - r\right), \nu_2 > 2r \quad \dots(14.15) \end{aligned}$$

**Aliter for (14.15).** (14.15) could also be obtained by substituting  $\frac{\nu_1}{\nu_2} F = \tan^2 \theta$  in (\*) and using the Beta integral :

$$\begin{aligned} 2 \int_0^{\pi/2} \sin^p \theta \cos^q \theta d\theta &= B\left(\frac{p+1}{2}, \frac{q+1}{2}\right) \\ \therefore \mu_r' &= \left(\frac{\nu_2}{\nu_1}\right)^r \cdot \frac{\Gamma[r + (\nu_1/2)] \Gamma[(\nu_2/2) - r]}{\Gamma(\nu_1/2) \Gamma(\nu_2/2)}, r < \frac{\nu_2}{2}. \quad \dots(14.16) \end{aligned}$$

In particular

$$\begin{aligned} \mu_1' &= \frac{\nu_2}{\nu_1} \cdot \frac{\Gamma[1 + (\nu_1/2)] \Gamma[(\nu_2/2) - 1]}{\Gamma(\nu_1/2) \Gamma(\nu_2/2)} \\ &= \frac{\nu_2}{\nu_2 - 2}, \nu_2 > 2 \quad [\because \Gamma(r) = (r-1) \Gamma(r-1)] \quad \dots[14.16(a)] \end{aligned}$$

Thus the mean of *F*-distribution is independent of  $\nu_1$ .

$$\begin{aligned} \mu_2' &= \left(\frac{\nu_2}{\nu_1}\right)^2 \cdot \frac{\Gamma[(\nu_1/2) + 2] \Gamma[(\nu_2/2) - 2]}{\Gamma(\nu_1/2) \Gamma(\nu_2/2)} \\ &= \left(\frac{\nu_2}{\nu_1}\right)^2 \cdot \frac{[(\nu_1/2) + 1] (\nu_1/2)}{[(\nu_2/2) - 1] [(\nu_2/2) - 2]} \\ &= \frac{\nu_2^2(\nu_1 + 2)}{\nu_1(\nu_2 - 2)(\nu_2 - 4)}, \nu_2 > 4. \\ \therefore \mu_2 &= \mu_2' - \mu_1'^2 = \frac{\nu_2^2(\nu_1 + 2)}{\nu_1(\nu_2 - 2)(\nu_2 - 4)} - \frac{\nu_2^2}{(\nu_2 - 2)^2} \\ &= \frac{2\nu_2^2(\nu_2 + \nu_1 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, \nu_2 > 4 \quad \dots[14.16(b)] \end{aligned}$$

Similarly, on putting  $r = 3$  and  $4$  in  $\mu_r'$ , we get  $\mu_3'$  and  $\mu_4'$  respectively, from which the central moments  $\mu_3$  and  $\mu_4$  can be obtained.

**Remark.** It has been proved that for large degrees of freedom,  $v_1$  and  $v_2$ ,  $F$  tends to  $N[1, 2 \{(1/v_1) + (1/v_2)\}]$  variate.

**14.5.3. Mode and Points of Inflexion of F-distribution.** We have

$$\log f(F) = C + \{(v_1/2) - 1\} \log F - \left(\frac{v_1 + v_2}{2}\right) \log \{1 + (v_1/v_2)F\}$$

where  $C$  is a constant independent of  $F$ .

$$\frac{\partial}{\partial F} [\log f(F)] = \left(\frac{v_1}{2} - 1\right) \cdot \frac{1}{F} - \frac{(v_1 + v_2)}{2} \cdot \frac{1}{\left[1 + \frac{v_1}{v_2}F\right]} \cdot \frac{v_1}{v_2}$$

$$f'(F) = \frac{\partial}{\partial F} f(F) = 0 \Rightarrow \frac{v_1 - 2}{2F} - \frac{v_1(v_1 + v_2)}{2(v_2 + v_1 F)} = 0$$

Hence  $F = \frac{v_2(v_1 - 2)}{v_1(v_2 + 2)}$  ... (14.17)

It can be easily verified that at this point  $f''(F) < 0$ . Hence

$$\text{Mode} = \frac{v_2(v_1 - 2)}{v_1(v_2 + 2)}$$

**Remarks 1.** Since  $F > 0$ , mode exists if and only if  $v_1 > 2$ .

2.  $\text{Mode} = \left(\frac{v_2}{v_2 + 2}\right) \cdot \left(\frac{v_1 - 2}{v_1}\right)$

Hence mode of  $F$ -distribution is always less than unity.

3. The points of inflexion of  $F$ -distribution exist for  $v_1 > 4$  and are equidistant from mode.

**Proof.** We have  $\frac{v_1}{v_2}F = \frac{X}{Y} \sim \beta_2(l, m)$ , (\*)

where  $l = v_1/2$  and  $m = v_2/2$ . We now find the points of inflexion of Beta distribution of second kind with parameters  $l$  and  $m$ .

If  $X \sim \beta_2(l, m)$ , its p.d.f. is

$$f(x) = \frac{1}{\beta(l, m)} \cdot \frac{x^{l-1}}{(1+x)^{l+m}} ; 0 \leq x < \infty \quad \dots (**)$$

Points of inflexion are the solution of

$$f''(x) = 0 \text{ and } f'''(x) \neq 0$$

From (\*\*),

$$\log f(x) = -\log \beta(l, m) + (l-1) \log x - (l+m) \log(1+x)$$

Differentiating twice w.r.t  $x$ , we get

$$\frac{f'(x)}{f(x)} = \frac{l-1}{x} - \frac{l+m}{1+x} \quad \dots (***)$$

$$\frac{f(x)f''(x) - [f'(x)]^2}{[f(x)]^2} = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2}$$

If  $f''(x) = 0$ , then we get

$$\begin{aligned} -\left[\frac{f'(x)}{f(x)}\right]^2 &= -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2} \\ \Rightarrow -\left[\frac{l-1}{x} - \frac{l+m}{1+x}\right]^2 &= -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2} \quad [\text{On using } (***)] \\ \Rightarrow \frac{l-1}{x^2}(l-1-1) - 2\frac{(l-1)(l+m)}{x(1+x)} + \frac{l+m}{(1+x)^2} \times (l+m+1) &= 0 \\ \Rightarrow (l-1)(l-2)(1+x)^2 - 2x(1+x)(l-1)(l+m) + x^2(l+m)(l+m+1) &= 0 \end{aligned} \quad \dots (****)$$

which is a quadratic in  $x$ . It can be easily verified that at these values of  $x$ ,  $f'''(x) \neq 0$ , if  $l > 2$ .

The roots of (\*\*\*\*) give the points of inflexion of  $\beta_2(l, m)$  distribution. The sum of the points of inflexion is equal to the sum of roots of (\*\*\*\*) and is given by :

$$\begin{aligned} &-\left[\frac{\text{Coefficient of } x \text{ in } (****)}{\text{Coefficient of } x^2 \text{ in } (****)}\right] \\ &= -\left[\frac{2(l-1)(l-2) - 2(l-1)(l+m)}{(l-1)(l-2) - 2(l-1)(l+m) + (l+m)(l+m+1)}\right] \\ &= \frac{2(l-1)[(l+m) - (l-2)]}{(l-1)(l-2) - (l-1)(l+m) - (l-1)(l+m) + (l+m)(l+m+1)} \\ &= \frac{2(l-1)(m+2)}{(l-1)[(l-2-l-m) + (l+m)[l+m+1-l+1]]} \\ &= \frac{2(l-1)(m+2)}{-(l-1)(m+2) + (l+m)(m+2)} \\ &= \frac{2(l-1)}{l+m-l+1} = \frac{2(l-1)}{(m+1)} \end{aligned}$$

$\therefore$  Sum of points of inflexion of  $\left(\frac{v_1}{v_2} F\right)$  distribution

$$= \frac{2(l-1)}{(m+1)} = \frac{2\left(\frac{v_1}{2} - 1\right)}{\left(\frac{v_2}{2} + 1\right)} = \frac{2(v_1-2)}{(v_2+2)}.$$

$\Rightarrow$  Sum of points of inflexion of  $F(v_1, v_2)$  distribution

$$\begin{aligned} &= \frac{v_2}{v_1} \cdot \frac{2(v_1-2)}{(v_2+2)}, \text{ provided } l = \frac{v_1}{2} > 2 \\ &= 2 \frac{v_2(v_1-2)}{v_1(v_2+2)} \\ &= 2 \text{ Mode, provided } v_1 > 4 \end{aligned}$$

Hence the points of inflexion of  $F(v_1, v_2)$  distribution, when they exist, (i.e., when  $v_1 > 4$ ), are equidistant from the mode.

4. Karl Pearson's coefficient of skewness is given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma} > 0,$$

since mean > 1 and mode < 1. Hence  $F$ -distribution is highly positively skewed.

5. The probability  $p(F)$  increases steadily at first until it reaches its peak (corresponding to the modal value which is less than 1) and then decreases slowly so as to become tangential at  $F = \infty$ , i.e.,  $F$ -axis is an asymptote to the right tail.

**Example 14-16.** When  $v_1 = 2$ , show that the significance level of  $F$  corresponding to a significant probability  $p$  is

$$F = \frac{v_2}{2} \left( p^{-(2/v_2)} - 1 \right)$$

where  $v_1$  and  $v_2$  have their usual meanings.

**Solution.** When  $v_1 = 2$ ,

$$\begin{aligned} dP(F) &= \frac{1}{B\left(1, \frac{v_2}{2}\right)} \cdot \frac{2}{v_2} \cdot \frac{dF}{\left[1 + \frac{2}{v_2} F\right]^{(v_2/2)+1}} && (\text{c.f. } \S 14-14a) \\ &= \frac{\Gamma(\frac{v_2}{2} + 1)}{\Gamma(1)\Gamma(v_2/2)} \times \frac{\frac{2}{v_2}}{\left(\frac{2}{v_2}\right)^{(v_2/2)+1} \left[F + \frac{v_2}{2}\right]^{(v_2/2)+1}} dF \\ &= \frac{\left[\frac{v_2}{2}\right]^{(v_2/2)+1}}{\left[F + \frac{v_2}{2}\right]^{(v_2/2)+1}} dF. \end{aligned}$$

$$\text{Hence } p = \int_F^\infty f(F) dF$$

$$= \left[\frac{v_2}{2}\right]^{(v_2/2)+1} \times \int_F^\infty \frac{dF}{\left[F + \frac{v_2}{2}\right]^{(v_2/2)+1}}$$

$$= \left[\frac{v_2}{2}\right]^{(v_2/2)+1} \times \left| \frac{\left[F + \frac{v_2}{2}\right]^{-(v_2/2)}}{-\frac{v_2}{2}} \right|_F^\infty$$

$$= \left[ \frac{\left(\frac{v_2}{2}\right)}{F + \frac{v_2}{2}} \right]^{v_2/2} = \frac{1}{\left[1 + \frac{2}{v_2} F\right]^{v_2/2}}$$

$$\Rightarrow p^{-2/v_2} = 1 + \frac{2F}{v_2} \Rightarrow F = \frac{v_2}{2} \left[ p^{-2/v_2} - 1 \right]$$

**Example 14-17.** If  $F(n_1, n_2)$  represent an *F*-variate with  $n_1$  and  $n_2$  d.f., prove that  $F(n_2, n_1)$  is distributed as  $1/F(n_1, n_2)$  variate. Deduce that

$$P[F(n_1, n_2) \geq c] = P\left[F(n_2, n_1) \leq \frac{1}{c}\right]$$

Or

Show how the probability points of  $F(n_2, n_1)$  can be obtained from those of  $F(n_1, n_2)$ .

**Solution.** Let  $X$  and  $Y$  be independent chi-square variates with  $n_1$  and  $n_2$  d.f. respectively. Then by definition, we have

$$F = \frac{(X/n_1)}{(Y/n_2)} \sim F(n_1, n_2)$$

$$\therefore \frac{1}{F} = \frac{(Y/n_2)}{(X/n_1)} \sim F(n_2, n_1) \quad \dots (*)$$

Hence the result.

We have :

$$\begin{aligned} P[F(n_1, n_2) \geq c] &= P\left[\frac{1}{F(n_1, n_2)} \leq \frac{1}{c}\right] \\ &= P\left[F(n_2, n_1) \leq \frac{1}{c}\right] \quad [\text{From } (*)] \end{aligned}$$

$$\text{Remark. } P[F(n_1, n_2) = c] = P\left[F(n_2, n_1) = \frac{1}{c}\right]$$

$$\text{Let } P[F(n_1, n_2) \geq c] = \alpha$$

i.e., let  $c$  be the upper  $\alpha$ -significant point of  $F(n_1, n_2)$  distribution.

$$\therefore 1 - \alpha = 1 - P[F(n_1, n_2) \geq c] = 1 - P\left[\frac{1}{F(n_1, n_2)} \leq \frac{1}{c}\right]$$

$$\Rightarrow \alpha = P\left[F(n_2, n_1) \leq \frac{1}{c}\right] = 1 - P\left[F(n_2, n_1) \geq \frac{1}{c}\right]$$

$$\Rightarrow P\left[F(n_2, n_1) \geq \frac{1}{c}\right] = 1 - \alpha$$

Thus  $(1 - \alpha)$  significant points of  $F(n_2, n_1)$  distribution are the reciprocal of  $\alpha$ -significant points of  $F(n_1, n_2)$  distribution, e.g.,

$$F_{8,4}(0.05) = 6.04 \Rightarrow F_{4,8}(0.95) = \frac{1}{6.04}$$

**Example 14-18.** Prove that if  $n_1 = n_2$ , the median of *F*-distribution is at  $F = 1$  and that the quartiles  $Q_1$  and  $Q_3$  satisfy the condition  $Q_1 Q_3 = 1$ .

**Solution.** Since  $n_1 = n_2 = n$ , (say), the median ( $M$ ) of  $F(n_1, n_2) = F(n, n)$  distribution is given by :

$$\begin{aligned} P[F(n, n) \leq M] &= 0.5 & \dots(*) \\ \Rightarrow P\left[\frac{1}{F(n, n)} \geq \frac{1}{M}\right] &= 0.5 \\ \Rightarrow P\left[F(n, n) \geq \frac{1}{M}\right] &= 0.5 & \left[ \because \frac{1}{F(m, n)} = F(n, m) \right] \\ \Rightarrow P\left[F(n, n) \leq \frac{1}{M}\right] &= 1 - P\left[F(n, n) \geq \frac{1}{M}\right] \\ &= 1 - 0.5 \\ &= 0.5 & \dots(**) \end{aligned}$$

From (\*) and (\*\*), we get

$$M = \frac{1}{M} \Rightarrow M^2 = 1 \Rightarrow M = 1$$

the negative value  $M = -1$ , is discarded since  $F > 0$ .

Hence the median of  $F(n, n)$  distribution is at  $F = 1$ .

Similarly, by definition of  $Q_1$  and  $Q_3$ , we have :

$$\begin{aligned} P[F(n, n) \leq Q_1] &= 0.25 & \dots(****) \\ \text{and } P[F(n, n) \geq Q_3] &= 0.25 \\ \Rightarrow P\left[\frac{1}{F(n, n)} \leq \frac{1}{Q_3}\right] &= 0.25 \\ \Rightarrow P\left[F(n, n) \leq \frac{1}{Q_3}\right] &= 0.25 & \left[ \because \frac{1}{F(m, n)} = F(n, m) \right] \dots(***) \end{aligned}$$

From (\*\*\*) and (\*\*\*\*), we get

$$Q_1 = \frac{1}{Q_3} \Rightarrow Q_1 Q_3 = 1$$

**Example 14.19.** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(0, 1)$ .

$$\text{Define } \bar{X}_k = \frac{1}{k} \sum_{i=1}^k X_i \text{ and } \bar{X}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^n X_i$$

Find the distribution of :

- (a)  $\frac{1}{2}(\bar{X}_k + \bar{X}_{n-k})$ ,
- (b)  $k\bar{X}_k^2 + (n-k)\bar{X}_{n-k}^2$
- (c)  $X_1^2/X_2^2$ ,
- (d)  $X_1/X_2$

[Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1989]

**Solution.** (a) Since  $X_1, X_2, \dots, X_n$  is a random sample from  $N(0, 1)$ ,

$$\bar{X}_k \sim N\left(0, \frac{1}{k}\right) \text{ and } \bar{X}_{n-k} \sim N\left(0, \frac{1}{n-k}\right) \dots(*)$$

Further, since  $(X_1, X_2, \dots, X_k)$  and  $(X_{k+1}, X_{k+2}, \dots, X_n)$  are independent,  $\bar{X}_k$  and  $\bar{X}_{n-k}$  are independent. Hence,

$$\begin{aligned} \frac{1}{2}(\bar{X}_k + \bar{X}_{n-k}) &= \frac{1}{2}\bar{X}_k + \frac{1}{2}\bar{X}_{n-k} \sim N\left(0, \frac{1}{4k} + \frac{1}{4(n-k)}\right) \\ \Rightarrow \quad \frac{1}{2}(\bar{X}_k + \bar{X}_{n-k}) &\sim N\left(0, \frac{n}{4k(n-k)}\right) \end{aligned}$$

(b) From (\*), we get

$$\begin{aligned} \frac{\bar{X}_k}{\sqrt{1/k}} &\sim N(0, 1) \quad \text{and} \quad \frac{\bar{X}_{n-k}}{\sqrt{1/(n-k)}} \sim N(0, 1) \\ \Rightarrow \quad k\bar{X}_k^2 &\sim \chi^2_{(1)} \quad \text{and} \quad (n-k)\bar{X}_{n-k}^2 \sim \chi^2_{(1)} \end{aligned}$$

Since  $\bar{X}_k$  and  $\bar{X}_{n-k}$  are independent, by additive property of chi-square distribution,

$$k\bar{X}_k^2 + (n-k)\bar{X}_{n-k}^2 \sim \chi^2_{(1+1)} = \chi^2_{(2)}$$

(c) Since  $X_1 \sim N(0, 1)$  and  $X_2 \sim N(0, 1)$  are independent,

$$X_1^2 \sim \chi^2_{(1)} \text{ and } X_2^2 \sim \chi^2_{(1)},$$

are also independent. Hence by definition of F-statistic,

$$\frac{X_1^2/1}{X_2^2/1} \sim F_{(1, 1)} \Rightarrow \frac{X_1^2}{X_2^2} \sim F_{(1, 1)}$$

(d)  $X_1/X_2$ , being the ratio of two independent standard normal variates is a standard Cauchy variate. [See Example 8-43].

### EXERCISE 14(e)

1. (a) Derive the distribution of  $F = S_1^2/S_2^2$ , where  $S_1^2$  and  $S_2^2$  are two independent unbiased estimates of the common population variance  $\sigma^2$ , defined by

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2; \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2$$

(b) Find the limiting form when the degrees of freedom of the  $\chi^2$  in the denominator tend to infinity and give an intuitive justification of the result.

2. (a) If  $X_1, X_2, \dots, X_m, X_{m+1}, \dots, X_{m+n}$  are independent normal variates with zero mean and standard deviation  $\sigma$ , obtain the distribution of

$$\sum_{i=1}^m X_i^2 \quad \text{and} \quad \sum_{i=m+1}^{m+n} X_i^2$$

Ans.  $F(m, n)$ .

(b) If  $X$  has an *F* distribution with  $n_1$  and  $n_2$  d.f., find the distribution of  $1/X$  and give one use of this result.

(c) If  $X$  is *t*-distributed, show that  $X^2$  is *F*-distributed.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1990*]

**Hint.** See § 14-5-6.

3. (a) Derive the distribution of the  $F$ -statistic on  $(n_1, n_2)$  degrees of freedom and show that the statistic  $\left(1 + \frac{n_1}{n_2} F\right)^{-1}$  has a Beta distribution.

(b) Show that the probability curve of the distribution of  $F$  is positively skewed.

4. Prove the following :

$$(i) \quad F_{n_1, n_2} = \frac{1}{F_{n_2, n_1}}$$

$$(ii) \quad F_{n_1, n_2} = \frac{n_2}{n_1} \cdot \frac{x}{1-x}, \text{ where } x \text{ has Beta-distribution.}$$

5. If  $X$  and  $Y$  are independent chi-square variates with  $v_1$  and  $v_2$  d.f. respectively, show that  $U = X + Y$  and  $V = \frac{v_2 X}{v_1 Y}$  are independently distributed.

Find the distribution of  $V$ .

6. Prove that if  $X$  has the  $F$ -distribution with  $(m, n)$  d.f. and  $Y$  has the  $F$ -distribution with  $(n, m)$  d.f., then for every  $a > 0$ ,

$$P(X \leq a) + P\left\{Y \leq \frac{1}{a}\right\} = 1$$

7. Show that the mode of the  $F$ -distribution with  $v_1 (\geq 2), v_2$  d.f. is given by  $\frac{v_2(v_1 - 2)}{v_1(v_2 + 2)}$  and is always less than unity.

8.  $X$  is  $F$ -variate with 2 and  $n (n \geq 2)$  degrees of freedom. Show that

$$P(F \geq k) = \left(1 + \frac{2k}{n}\right)^{-n/2}$$

[Gujarat Univ. B.Sc., 1992]

Deduce the significance level of  $F$  corresponding to the significance level of probability  $P$ .

9. Let  $X_1, X_2$  be independent random variables following the density law  $f(x) = e^{-x}, 0 < x < \infty$ . Show that

$Z = X_1/X_2$ , has an  $F$ -distribution.

10. (a) If  $\bar{X} \sim F(n_1, n_2)$ , show that its mean is independent of  $n_1$ .

(b) Obtain the mode of  $F$ -distribution with  $(n_1, n_2)$  d.f. and show that it lies between 0 and 1.

(c) Show that for  $F$ -distribution with  $n_1$  and  $n_2$  d.f., the points of inflexion exist if  $n_1 > 4$  and are equidistant from the mode.

11.  $X$  is a binomial variate with parameters  $n$  and  $p$  and  $F_{v_1, v_2}$  is an  $F$ -statistic with  $v_1$  and  $v_2$  d.f. Prove that

$$P(X \leq k-1) = P\left[F_{2k, 2(n-k+1)} > \frac{n-k+1}{k} \cdot \frac{p}{1-p}\right]$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1985]

Hint. If  $X \sim B(n, p)$ , then we have [c.f. Example 7.23]

$$\begin{aligned}
 P(X \leq k-1) &= (n-k+1) \cdot \binom{n}{k-1} \int_0^q t^{n-k} (1-t)^{k-1} dt \\
 &= \frac{1}{B(k, n-k+1)} \int_0^q t^{n-k} (1-t)^{k-1} dt \\
 P\left[F_{2k, 2(n-k+1)} > \frac{n-k+1}{k} \left(\frac{p}{1-p}\right)\right] &= \int_{\frac{n-k+1}{k} \cdot \frac{p}{q}}^{\infty} p[F_{2k, 2(n-k+1)}] dF \\
 &= \frac{1}{B(k, n-k+1)} \int_{\frac{n-k+1}{k} \cdot \frac{p}{q}}^{\infty} \frac{[k/(n-k+1)]^k \cdot F^{k-1} dF}{\left[1 + \frac{kF}{n-k+1}\right]^{n+1}} \\
 &= \frac{1}{B(k, n-k+1)} \int_0^q y^{n-k} (1-y)^{k-1} dy
 \end{aligned}$$

where  $1 + \frac{kF}{n-k+1} = \frac{1}{y}$ .

12. (a) If  $X \sim F(n_1, n_2)$  distribution, show that

$$U = \frac{n_1 X}{n_2 + n_1 X} \sim \beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$$

[Delhi Univ. B.Sc. (Maths. Hons.), 1992]

Hence obtain the distribution function of  $X$ .

Hint. The distribution function of  $X \sim F(n_1, n_2)$  is given by

$$\begin{aligned}
 G_X(x) &= \int_0^x f(F) dF = \int_0^y h(u) du, \quad \left[y = \frac{n_1 x}{n_2 + n_1 x}\right] \\
 &= \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \int_0^y u^{\frac{n_1}{2}-1} (1-u)^{\frac{n_2}{2}-1} du \\
 &= I_y\left(\frac{n_1}{2}, \frac{n_2}{2}\right),
 \end{aligned}$$

where  $I_x(p, q) = \frac{1}{B(p, q)} \int_0^x t^{p-1} (1-t)^{q-1} dt$ ,

is the incomplete Beta function. Hence the distribution function of  $F$  distribution can be obtained from the tables of incomplete Beta function.

(b)  $X \sim F(m, n)$ , show that

$$W = \frac{m X / n}{1 + (m X / n)} \sim \beta_1\left(\frac{1}{2} m, \frac{1}{2} n\right)$$

Deduce the variance of  $X$  from p.d.f. of  $W$ .

[Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1989]

13. Let  $X_1$  and  $X_2$  be a random sample of size 2 from  $N(0, 1)$  and  $Y_1$  and  $Y_2$  be a random sample of size 2 from  $N(1, 1)$ , and let the  $Y_i$ 's be independent of the  $X_i$ 's. Find the distribution of the following :

- (i)  $(X_1 - X_2)/\sqrt{2}$
- (ii)  $(X_1 + X_2)^2/(X_2 - X_1)^2$
- (iii)  $\bar{X} + \bar{Y}$
- (iv)  $(Y_1 + Y_2 - 2)^2/(X_2 - X_1)^2$
- v)  $(X_1 + X_2)/\sqrt{[(X_2 - X_1)^2 + (Y_2 - Y_1)^2]/2}$
- vi)  $[(Y_1 - Y_2)^2 + (X_1 - X_2)^2 + (X_1 + X_2)^2]/2$

[Delhi Univ. B.Sc. (Maths. Hons.), 1988, 1987]

- Ans.** (i)  $N(0, 1)$ , (ii)  $F(1, 1)$ , (iii)  $N(1, 1)$   
 (iv)  $F(1, 1)$ , (v)  $t_{(2)}$ , (vi)  $\chi^2_{(3)}$ .

14. Let  $X_i \sim N(i, i^2)$ ,  $i = 1, 2, 3$  be independent random variables. Using only the three random variables  $X_1$ ,  $X_2$ , and  $X_3$ , give an example of a statistic that has :

- (i) A chi-square distribution with 3 d.f.
- (ii) An  $F$ -distribution with (1, 2) d.f.
- (iii) A  $t$ -distribution with 2 d.f.

[Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1986]

- Ans.** Hint.  $Z_i = (X_i - i)/i$ ,  $i = 1, 2, 3$ , are i.i.d.  $N(0, 1)$ .

$$(i) \sum_{i=1}^3 Z_i^2 \sim \chi^2_{(3)}; (ii) \frac{Z_1^2}{Z_2^2 + Z_3^2} \sim F(1, 2); (iii) \frac{Z_1}{[(Z_2^2 + Z_3^2)/2]^{1/2}} \sim t_{(2)}$$

15. Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$ . Define :

$$\bar{X}_k = \frac{1}{k} \sum_{i=1}^k X_i, \quad \bar{X}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^n X_i, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$S_k^2 = \frac{1}{k-1} \sum_{i=1}^k (X_i - \bar{X}_k)^2, \quad S_{n-k}^2 = \frac{1}{n-k-1} \sum_{i=k+1}^n (X_i - \bar{X}_{n-k})^2$$

and  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ .

Answer the following questions :

- (i) What is the distribution of

$$\sigma^{-2} [(k-1)S_k^2 + (n-k-1)S_{n-k}^2]?$$

- (ii) What is the distribution of  $S_k^2/S_{n-k}^2$ ?

- (iii) What is the distribution of  $(\bar{X} - \mu) \sqrt{n}/S$ ?

[Delhi Univ. B.Sc. (Maths Hons.), 1989]

- (iv) What is the distribution of  $\frac{1}{2}(\bar{X}_k + \bar{X}_{n-k})$ ?

- (v) What is the distribution of  $(X_i - \mu)^2/\sigma^2$ ?

- Ans.** (i)  $\chi^2_{(k-1)+(n-k-1)} = \chi^2_{(n-2)}$ ; (ii)  $F_{(k-1, n-k-1)}$

(iii)  $t_{(n-1)}$ ; (iv)  $N\left(\mu, \frac{n\sigma^2}{4k(n-k)}\right)$ , (v)  $\chi^2_{(1)}$

16. If  $X \sim F(1, n)$ , show that

$$\left(n - \frac{1}{2}\right) \log [1 + (X/n)] \sim \chi^2_{(1)}$$

for large  $n$ .

17. If  $X_1, X_2, X_3$  and  $X_4$  are independent observations from  $N(0, 1)$  population, state giving reasons, the sampling distributions of

$$(i) U = \frac{\sqrt{2} X_3}{\sqrt{X_1^2 + X_2^2}} \quad \text{and} \quad (ii) V = \frac{3X_4^2}{X_1^2 + X_2^2 + X_3^2}$$

Ans. (i)  $U \sim t_{(2)}$ ; (ii)  $V \sim F(1, 3)$ .

18. Let  $(X_1, X_2)$  be a random sample from  $N(0, 1)$ . Answer the following, giving reasons :

- (i) What is the distribution of  $(X_2 - X_1)^2/2$ ?
- (ii) What is the distribution of  $(X_1 + X_2)^2/(X_2 - X_1)^2$ ?
- (iii) What is the distribution of  $(X_2 + X_1)/\sqrt{(X_1 - X_2)^2}$ ?
- (iv) What is the distribution of  $1/Z$ , if  $Z = X_1^2/X_2^2$ ?

[Delhi Univ. B.Sc. (Maths. Hons.), 1992]

Ans. (i)  $\chi^2_{(1)}$ ; (ii)  $F(1, 1)$ ; (iii) Standard Cauchy; (iv)  $F(1, 1)$

**14-5-4. Applications of F-distribution.** F-distribution has the following applications in Statistical theory.

**14-5-5. F-test for Equality of Population Variances.** Suppose we want to test (i) whether two independent samples  $x_i$ , ( $i = 1, 2, \dots, n_1$ ) and  $y_j$ , ( $j = 1, 2, \dots, n_2$ ) have been drawn from the normal populations with the same variance  $\sigma^2$ , (say), or (ii) whether the two independent estimates of the population variance are homogeneous or not.

Under the null hypothesis ( $H_0$ ) that (i)  $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ , i.e., the population variances are equal or (ii) Two independent estimates of the population variance are homogeneous, the statistic  $F$  is given by

$$F = \frac{S_X^2}{S_Y^2} \quad \dots (14-18)$$

where  $S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$  and  $S_Y^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \dots (14-18a)$

are unbiased estimates of the common population variance  $\sigma^2$  obtained from two independent samples and it follows Snedecor's F-distribution with  $(v_1, v_2)$  d.f. [where  $v_1 = n_1 - 1$  and  $v_2 = n_2 - 1$ ].

$$\begin{aligned} \text{Proof. } F &= \frac{S_X^2}{S_Y^2} = \left[ \frac{n_1}{n_1 - 1} S_X^2 \right] / \left[ \frac{n_2}{n_2 - 1} S_Y^2 \right] \\ &= \left[ \frac{n_1 S_X^2}{\sigma_X^2} \cdot \frac{1}{(n_1 - 1)} \right] / \left[ \frac{n_2 S_Y^2}{\sigma_Y^2} \cdot \frac{1}{(n_2 - 1)} \right] \\ &\quad (\because \sigma_X^2 = \sigma_Y^2 = \sigma^2 \text{ under } H_0) \end{aligned}$$

Since  $\frac{n_1 s_x^2}{\sigma_x^2}$  and  $\frac{n_2 s_y^2}{\sigma_y^2}$  are independent chi-square variates with  $(n_1 - 1)$  and  $(n_2 - 1)$  d.f. respectively,  $F$  follows Snedecor's  $F$ -distribution with  $(n_1 - 1, n_2 - 1)$  d.f. (c.f. § 14-5).

**Remarks 1.** In (14-18), greater of the two variances  $s_x^2$  and  $s_y^2$  is to be taken in the numerator and  $n_1$  corresponds to the greater variance.

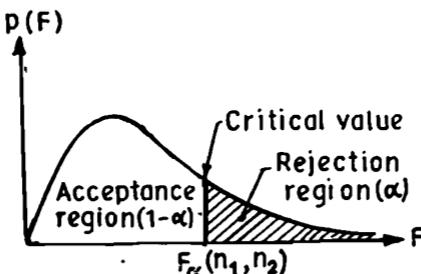
By comparing the calculated value of  $F$  obtained by using (14-18) for the two given samples with the tabulated value of  $F$  for  $(n_1, n_2)$  d.f. at certain level of significance (5% or 1%),  $H_0$  is either rejected or accepted.

**2. Critical values of  $F$ -distribution.** The available  $F$ -tables (given in the Appendix at the end of the book) 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 significant value  $F_\alpha (n_1, n_2)$  at level of significance  $\alpha$  and  $(n_1, n_2)$  d.f. is determined by

$$P[F > F_\alpha (n_1, n_2)] = \alpha, \quad \dots (*)$$

as shown in the following diagram.

CRITICAL VALUES OF F-DISTRIBUTION



From Remark to Example 14-17, we have the following reciprocal relation between the upper and lower  $\alpha$ -significant points of  $F$ -distribution :

$$\begin{aligned} F_\alpha (n_1, n_2) &= \frac{1}{F_{1-\alpha} (n_2, n_1)} \\ \Rightarrow F_\alpha (n_1, n_2) \times F_{1-\alpha} (n_2, n_1) &= 1 \end{aligned} \quad \dots (**)$$

The critical values of  $F$  for left tail test  $H_0 : \sigma_1^2 = \sigma_2^2$  against  $H_1 : \sigma_1^2 < \sigma_2^2$  are given by  $F < F_{n_1-1, n_2-1}(1-\alpha)$ , and for the two tailed test,  $H_0 : \sigma_1^2 = \sigma_2^2$  against  $H_1 : \sigma_1^2 \neq \sigma_2^2$  are given by  $F > F_{n_1-1, n_2-1}(\alpha/2)$  and  $F < F_{n_1-1, n_2-1}(1-\alpha/2)$  [For details, see § 16-7-5].

**Example 14-20.** Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins show the sample standard deviations of their weights as 0.8 and 0.5 respectively. Assuming that the weight distributions are normal, test the hypothesis that the true variances are equal, against the alternative that they are not, at the 10% level. [Assume that  $P(F_{10, 8} \geq 3.35) = 0.05$  and  $P(F_{8, 10} \geq 3.07) = 0.05$ ].

**Solution.** We want to test Null Hypothesis,  $H_0 : \sigma_x^2 = \sigma_y^2$ , against the Alternative Hypothesis,  $H_1 : \sigma_x^2 \neq \sigma_y^2$  (Two-tailed).

We are given :

$$n_1 = 11, n_2 = 9, s_x = 0.8 \text{ and } s_y = 0.5.$$

Under the null hypothesis,  $H_0: \sigma_x^2 = \sigma_y^2$ , the statistic

$$F = \frac{s_x^2}{s_y^2}$$

follows *F*-distribution with  $(n_1 - 1, n_2 - 1)$  d.f.

Now

$$n_1 s_x^2 = (n_1 - 1) S_x^2$$

$$\therefore S_x^2 = \left( \frac{n_1}{n_1 - 1} \right) s_x^2 = \left( \frac{11}{10} \right) \times (0.8)^2 = 0.704$$

$$\text{Similarly, } S_y^2 = \left( \frac{n_2}{n_2 - 1} \right) s_y^2 = \left( \frac{9}{8} \right) \times (0.5)^2 = 0.28125$$

$$\therefore F = \frac{0.704}{0.28125} = 2.5$$

The significant values of *F* for two tailed test at level of significance  $\alpha = 0.10$  are :

$$\begin{aligned} F &> F_{10,8}(\alpha/2) = F_{10,8}(0.05) \\ \text{and } F &< F_{10,8}(1 - \alpha/2) = F_{10,8}(0.95) \end{aligned} \quad \dots (*)$$

We are given the tabulated (significant) values :

$$P[F_{10,8} \geq 3.35] = 0.05 \Rightarrow F_{10,8}(0.05) = 3.35 \quad \dots (**)$$

$$\text{Also } P[F_{8,10} \geq 3.07] = 0.05 \Rightarrow P\left[\frac{1}{F_{8,10}} \leq \frac{1}{3.07}\right] = 0.05$$

$$\Rightarrow P[F_{10,8} \leq 0.326] = 0.05 \Rightarrow P[F_{10,8} \geq 0.326] = 0.95 \quad \dots (***)$$

Hence from (\*), (\*\*) and (\*\*\*) , the critical values for testing  $H_0: \sigma_x^2 = \sigma_y^2$ , against  $H_1: \sigma_x^2 \neq \sigma_y^2$  at level of significance  $\alpha = 0.10$  are given by :

$$F > 3.35 \text{ and } F < 0.326 = 0.33$$

Since, the calculated value of *F* (=2.5) lies between 0.33 and 3.35, it is not significant and hence null hypothesis of equality of population variances may be accepted at level of significance  $\alpha = 0.10$ .

**Example 14-21.** In one sample of 8 observations, the sum of the squares of deviations of the sample values from the sample mean was 84.4 and in the other sample of 10 observations it was 102.6. Test whether this difference is significant at 5 per cent level, given that the 5 per cent point of *F* for  $n_1 = 7$  and  $n_2 = 9$  degrees of freedom is 3.29. [Delhi Univ. B.Sc. (Maths Hons.), 1986]

**Solution.** Here  $n_1 = 8, n_2 = 10$

$$\text{and } \sum(x - \bar{x})^2 = 84.4, \quad \sum(y - \bar{y})^2 = 102.6$$

$$\therefore S_x^2 = \frac{1}{n_1 - 1} \sum(x - \bar{x})^2 = \frac{84.4}{7} = 12.057$$

$$S_y^2 = \frac{1}{n_2 - 1} \sum(y - \bar{y})^2 = \frac{102.6}{9} = 11.4$$

Under  $H_0 : \sigma_x^2 = \sigma_y^2 = \sigma^2$ , i.e., the estimates of  $\sigma^2$  given by the samples are homogeneous, the test statistic is

$$F = \frac{S_x^2}{S_y^2} = \frac{12.057}{11.4} = 1.057$$

Tabulated  $F_{0.05}$  for (7, 9) d.f. is 3.29.

Since calculated  $F < F_{0.05}$ ,  $H_0$  may be accepted at 5% level of significance.

**Example 14.22.** Two random samples gave the following results :

Sample	Size	Sample mean	Sum of squares of deviations from the mean
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population at 5% level of significance.

[Given :  $F_{0.05}(9, 11) = 2.90$ ,  $F_{0.05}(11, 9) = 3.10$  (approx.)

and  $t_{0.05}(20) = 2.086$ ,  $t_{0.05}(22) = 2.07$ ]

[Delhi Univ. MCA, 1987]

**Solution.** A normal population has two parameters, viz., mean  $\mu$  and variance  $\sigma^2$ . To test if two independent samples have been drawn from the same normal population we have to test (i) the equality of population means, and (ii) the equality of population variances.

*Null Hypothesis* : The two samples have been drawn from the same normal population, i.e.,  $H_0 : \mu_1 = \mu_2$  and  $\sigma_1^2 = \sigma_2^2$ .

Equality of means will be tested by applying  $t$ -test and equality of variances will be tested by applying  $F$ -test. Since  $t$ -test assumes  $\sigma_1^2 = \sigma_2^2$ , we shall first apply  $F$ -test and then  $t$ -test.

We are given  $n_1 = 10$ ,  $n_2 = 12$ ;  $\bar{x}_1 = 15$ ,  $\bar{x}_2 = 14$   
 $\sum(x_1 - \bar{x}_1)^2 = 90$ ,  $\sum(x_2 - \bar{x}_2)^2 = 108$

### F-test

$$\text{Here } S_1^2 = \frac{1}{n_1 - 1} \sum(x_1 - \bar{x}_1)^2 = \frac{90}{9} = 10$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum(x_2 - \bar{x}_2)^2 = \frac{108}{11} = 9.82$$

Since  $S_1^2 > S_2^2$ , under  $H_0 : \sigma_1^2 = \sigma_2^2$ , the test statistic is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) = F(9, 11)$$

$$\text{Now } F = \frac{10}{9.82} = 1.018$$

Tabulated  $F_{0.05}(9, 11) = 2.90$

Since calculated  $F$  is less than tabulated  $F$  it is not significant. Hence null hypothesis of equality of population variances may be accepted.

Since  $\sigma_1^2 = \sigma_2^2$ , we can now apply *t* test for testing  $H_0 : \mu_1 = \mu_2$ .

**t-test.** Under  $H_0' : \mu_1 = \mu_2$ , against alternative hypothesis,  $H_1' : \mu_1 \neq \mu_2$ , the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1 + n_2 - 2} = t_{20}$$

$$\text{where } S^2 = \frac{1}{n_1 + n_2 - 2} [\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2] \\ = \frac{1}{20} [90 + 108] = 9.9$$

$$\therefore t = \frac{15 - 14}{\sqrt{9.9 \left( \frac{1}{10} + \frac{1}{12} \right)}} = \frac{1}{\sqrt{9.9 \times \frac{11}{60}}} \\ = \frac{1}{\sqrt{1.815}} = 0.742$$

Now  $t_{0.05}$  for 20 d.f. = 2.086

Since  $|t| < t_{0.05}$ , it is not significant. Hence the hypothesis  $H_0' : \mu_1 = \mu_2$  may be accepted. Since both the hypotheses, i.e.,  $H_0' : \mu_1 = \mu_2$  and  $H_0 : \sigma_1^2 = \sigma_2^2$  are accepted, we may regard that the given samples have been drawn from the same normal population.

### EXERCISE 14(f)

1. (a) If  $\chi_1^2$  and  $\chi_2^2$  are independent chi-square variates with  $n_1$  and  $n_2$  d.f., obtain the probability density function of *F*-statistic defined by

$$F = \frac{(\chi_1^2/n_1)}{(\chi_2^2/n_2)}$$

Mention the types of hypotheses which are tested with the help of this statistic.

(b) Explain why the larger variance is placed in the numerator of the statistic *F*. Discuss the application of *F*-test in testing if two variances are homogeneous.

2. An investigator, newly appointed, was made to take ten independent measurements on the maximum internal diameter of a pot at specified equal intervals of time and the standard deviation of these ten observations was found

to be 0.0345 mm. After he had been some time on similar jobs, he was asked to repeat this experiment an equal number of times and the standard deviation of the new set of ten observations was found to be 0.0285 mm. Can it be concluded that the investigator has become more consistent (*i.e.* less variable) with practice?

3. (a) Two independent samples of 8 and 7 items respectively had the following values of the variables.

Sample I :	9	11	13	11	15	9	12	14
Sample II :	10	12	10	14	9	8	10	

Do the estimates of population variance differ significantly?

[Delhi Univ. B.Sc., 1992]

(b) Five measurements of the output of two units have given the following results (in kilograms of material per one hour of operation).

Unit A :	14.1	10.1	14.7	13.7	14.0
Unit B :	14.0	14.5	13.7	12.7	14.1

Assuming that both samples have been obtained from normal populations, test at 10% significance level if the two populations have the same variance, it being given that  $F_{0.95}(4, 4) = 6.39$

[Calcutta Univ. B.Sc. (Maths. Hons.), 1991]

(c) In one sample of 10 observations from a normal population, the sum of the squares of the deviations of the sample values from the sample mean is 102.4 and in another sample of 12 observations from another normal population, the sum of the squares of the deviations of the sample values from the sample mean is 120.5. Examine whether the two normal populations have the same variance.

4. (a) Two random samples of sizes 8 and 11, drawn from two normal populations, are characterised as follows:

Population from which the sample is drawn	Size of sample	Sum of observations	Sum of squares of observations
I	8	9.6	61.52
II	11	16.5	73.26

You are to decide if the two populations can be taken to have the same variance. What test function would you use? How is it distributed and what value it has in this sampling experiment?

(b) 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,

Ans.  $H_0: \sigma_1^2 = \sigma_2^2$  (both engineers are equally consistent).  $F = 2.4$ . Not significant.

(c) 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

Can it be said that the two samples come from the same normal population?

**Ans.**  $H_0: \mu_1 = \mu_2; t = 1.9$ , Not significant.

$H'_0: \sigma_1^2 = \sigma_2^2, F = 4.08 < 6.26 [F_{0.05}(5, 4)]$ . Not significant.

Hence the two samples have come from the same normal population.

5. (a) Two random samples drawn from two normal populations are :

*Sample I* : 20, 16, 26, 27, 23, 22, 18, 24, 25, 19

*Sample II* : 27, 33, 42, 35, 32, 34, 38, 28, 41, 43, 30, 37

Obtain estimates of the variances of the populations and test whether the populations have same variances.

[Given  $F_{0.05} = 3.11$  for 11 and 9 degrees of freedom.]

(b) Test  $H_0: \sigma_1^2 = \sigma_2^2$  against  $H_1: \sigma_1^2 \neq \sigma_2^2$

given  $n_1 = 25, \sum (x_i - \bar{x})^2 = 164 \times 24,$

$n_2 = 21, \sum (y_j - \bar{y})^2 = 190 \times 21.$

Make necessary assumptions, stating them.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1987*]

(c) The diameters of two random samples, each of size 10, of bullets produced by two machines have standard deviations  $s_1 = 0.01$  and  $s_2 = 0.015$ . Assuming that the diameters have independent distributions  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$ , test the hypothesis that, the two machines are equally good by testing :

$H_0: \sigma_1 = \sigma_2$  against  $H_1: \sigma_1 \neq \sigma_2.$

6. The following table shows the yield of corn in bushels per plot in 20 plots, half of which are treated with phosphate as fertiliser.

Treated	5	0	8	3	6	1	0	3	3	1
Untreated	1	4	1	2	3	2	5	0	2	0

Test whether the treatment by phosphate has

(i) changed the variability of the plot yields,

(ii) improved the average yield of corn.

7. (a) The following figures give the prices in rupees of a certain commodity in a sample of 15 shops selected at random from a city A and those in a sample of 13 shops from another city B.

City A :	7.41	7.77	7.44	7.40	7.38	7.93	7.58
	8.28	7.23	7.52	7.82	7.71	7.84	7.63
City B	7.08	7.49	7.42	7.04	6.92	7.22	7.68
	7.24	7.74	7.81	7.28	7.43	7.47	

Assuming that the distribution of prices in the two cities is normal, answer the following :

(i) Is it possible that the average price of city B is Rs. 7.20?

(ii) Is the observed variance in the first sample consistent with the hypothesis that the standard deviation of prices in city A is Rs. 0.30?

(iii) Is it reasonable to say that the variability of prices in the two cities is the same?

(iv) Is it reasonable to say that the average prices are the same in two cities?

**14.5.6. Relation between  $t$  and  $F$  distributions.** In  $F$ -distribution with  $(v_1, v_2)$  d.f. [c.f. 14.5 (a)], take  $v_1 = 1$ ,  $v_2 = v$  and  $t^2 = F$ , i.e.,  $dF = 2t dt$ . Thus the probability differential of  $F$  transforms to

$$\begin{aligned} dG(t) &= \frac{(1/v)^{1/2}}{B\left(\frac{1}{2}, \frac{v}{2}\right)} \cdot \frac{(t^2)^{\frac{1}{2}-1}}{\left[1 + \frac{t^2}{v}\right]^{(v+1)/2}} 2t dt, \quad 0 \leq t^2 < \infty \\ &= \frac{1}{\sqrt{v} B\left(\frac{1}{2}, \frac{v}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{v}\right]^{(v+1)/2}} dt, \quad -\infty < t < \infty \end{aligned}$$

the factor 2 disappearing since the total probability in the range  $(-\infty, \infty)$  is unity. This is the probability function of Student's  $t$ -distribution with  $v$  d.f. Hence we have the following relation between  $t$  and  $F$  distributions.

'If a statistic  $t$  follows Student's  $t$ -distribution with  $n$  d.f., then  $t^2$  follows Snedecor's  $F$ -distribution with  $(1, n)$  d.f. Symbolically,

$$\left. \begin{array}{l} \text{if} \quad t \sim t_{(n)} \\ \text{then} \quad t^2 \sim F_{(1, n)} \end{array} \right\} \quad \dots(14.19)$$

**Aliter Proof of (14.19).** If  $\xi \sim N(0, 1)$  and  $X \sim \chi^2_{(n)}$  are independent r.v.'s then :

$$U = \xi^2 \sim \chi^2_{(1)} \quad [\text{Square of a S.N.V.}]$$

$$\text{and} \quad t = \frac{\xi}{\sqrt{X/n}} \sim t_{(n)}$$

$$\Rightarrow t^2 = \frac{\xi^2}{(X/n)} = \frac{(\xi^2/1)}{(X/n)},$$

being the ratio of two independent chi-square variates divided by their respective degrees of freedom is  $F(1, n)$  variate.

$$\text{Hence} \quad t^2 \sim F(1, n)$$

With the help of relation (14.19), all the uses of  $t$ -distribution can be regarded as the applications of  $F$ -distribution also, e.g., for test for a single mean, instead of computing

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

we may compute

$$F = t^2 = \frac{n(\bar{x} - \mu)^2}{S^2}$$

and then apply  $F$ -test with  $(1, n)$  d.f. and so on.

Similarly, we can write the test statistic  $F$  from § 14.2.9, § 14.2.10 and § 14.2.11 for testing the significance of an observed sample correlation coefficient, regression coefficient and partial correlation coefficient respectively.

**Example 14.23.** Given :  $P[F(10, 12) > 2.753] = 0.05$   
 $= P[F(1, 12) > 4.747]$

find  $P[F(12, 10) > (2.753)^{-1}]$ , and  $P[-\sqrt{4.747} < t_{12} < \sqrt{4.747}]$

**Solution.**

$$\begin{aligned} P[F(12, 10) > (2.753)^{-1}] &= P\left[\frac{1}{F(12, 10)} < 2.753\right] \\ &= P[F(10, 12) < 2.753] \\ &= 1 - P[F(10, 12) > 2.753] \\ &= 1 - 0.05 = 0.95 \end{aligned}$$

$$\begin{aligned} P[-\sqrt{4.747} < t_{12} < \sqrt{4.747}] &= P(t_{12}^2 < 4.747) \\ &= P[F(1, 12) < 4.747] \\ &= 1 - P[F(1, 12) > 4.747] \\ &= 1 - 0.05 = 0.95 \end{aligned}$$

**14.5.7. Relation between *F* and  $\chi^2$ .** In  $F(n_1, n_2)$  distribution if we let  $n_2 \rightarrow \infty$ , then  $\chi^2 = n_1 F$  follows  $\chi^2$ -distribution with  $n_1$  d.f.

**Proof.** We have

$$p(F) = \frac{(n_1/n_2)^{n_1/2} F^{(n_1/2)-1}}{\Gamma(n_1/2) \Gamma(n_2/2)} \cdot \left[ 1 + \frac{n_1}{n_2} F \right]^{(n_1+n_2)/2}, \quad 0 < F < \infty$$

In the limit as  $n_2 \rightarrow \infty$ , we have

$$\frac{\Gamma(n_1+n_2)/2}{n_2^{n_1/2} \Gamma(n_2/2)} \rightarrow \frac{(n_2/2)^{n_1/2}}{n_2^{n_1/2}} = \frac{1}{2^{n_1/2}}$$

$$\left[ \therefore \frac{\Gamma(n+k)}{\Gamma(n)} \rightarrow n^k \text{ as } n \rightarrow \infty. \text{ (c.f. Remark below.)} \right]$$

$$\begin{aligned} \text{Also } \lim_{n_2 \rightarrow \infty} \left[ 1 + \frac{n_1}{n_2} F \right]^{(n_1+n_2)/2} &= \lim_{n_2 \rightarrow \infty} \left[ \left( 1 + \frac{n_1}{n_2} F \right)^{n_2} \right]^{1/2} \\ &\times \lim_{n_2 \rightarrow \infty} \left( 1 + \frac{n_1}{n_2} F \right)^{n_1/2} \\ &= \exp(n_1 F/2) = \exp(\chi^2/2) \quad (\because n_1 F = \chi^2) \end{aligned}$$

Hence in the limit, the p.d.f. of  $\chi^2 = n_1 F$  becomes

$$\begin{aligned} dP(\chi^2) &= \frac{(n_1/2)^{n_1/2} e^{-\chi^2/2}}{\Gamma(n_1/2)} \cdot \left( \frac{\chi^2}{n_1} \right)^{(n_1/2)-1} d\left( \frac{\chi^2}{n_1} \right) \\ &= \frac{1}{2^{n_1/2} \Gamma(n_1/2)} \cdot e^{-\chi^2/2} (\chi^2)^{(n_1/2)-1} d\chi^2, \quad 0 < \chi^2 < \infty \end{aligned}$$

which is the p.d.f. of chi-square distribution with  $n_1$  d.f.

**Remark.**  $\lim_{n \rightarrow \infty} \frac{\Gamma(n+k)}{\Gamma(n)} = \lim_{n \rightarrow \infty} \frac{(n+k-1)!}{(n-1)!}$ , (for large  $n$ )

$$\approx \lim_{n \rightarrow \infty} \frac{\sqrt{2\pi} e^{-(n+k-1)} (n+k-1)^{n+k-1/2}}{\sqrt{2\pi} e^{-(n-1)} (n-1)^{n-1/2}}$$

(On using Stirling's approximation for  $n!$  as  $n \rightarrow \infty$ .)

$$\begin{aligned} &= e^{-k} \lim_{n \rightarrow \infty} \frac{n^{n+k-\frac{1}{2}} \left(1 + \frac{k-1}{n}\right)^{n+k-\frac{1}{2}}}{n^{n-\frac{1}{2}} \left(1 - \frac{1}{n}\right)^{n-\frac{1}{2}}} \\ &= e^{-k} n^k \frac{\lim_{n \rightarrow \infty} \left(1 + \frac{k-1}{n}\right)^n \lim_{n \rightarrow \infty} \left(1 + \frac{k-1}{n}\right)^{k-\frac{1}{2}}}{\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^{-\frac{1}{2}}} \\ &= e^{-k} n^k \left[ \frac{e^{(k-1)}}{e^{-1} \cdot 1} \right] = n^k \\ &\quad \lim_{n \rightarrow \infty} \frac{\Gamma(n+k)}{\Gamma n} \approx n^k \end{aligned}$$

**14.5.8. F-test for Testing the Significance of an Observed Multiple Correlation Coefficient.** If  $R$  is the observed multiple correlation coefficient of a variate with  $k$  other variates in a random sample of size  $n$  from a  $(k+1)$  variate population, then Prof. R.A. Fisher proved that under the null hypothesis ( $H_0$ ) that the multiple correlation coefficient in the population is zero, the statistic

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}. \quad \dots(14.20)$$

conforms to  $F$ -distribution with  $(k, n-k-1)$  d.f.

**14.5.9. F-test for significance of an observed sample correlation Ratio  $\eta_{yx}$ .** Under the null hypothesis that population correlation ratio is zero, the test statistic is

$$F = \frac{\eta^2}{1-\eta^2} \cdot \frac{N-h}{h-1} \sim F(h-1, N-h) \quad \dots(14.21)$$

where  $N$  is the size of the sample (from a bi-variate normal population) arranged in  $h$  arrays.

**14.5.10. F-test for Testing the Linearity of Regression.** For a sample of size  $N$  arranged in  $h$  arrays, from a bi-variate normal population, the test statistic for testing the hypothesis of linearity of regression is,

$$F = \frac{\eta^2 - r^2}{1-\eta^2} \cdot \frac{N-h}{h-2} \sim F(h-2, N-h) \quad \dots(14.22)$$

**14.5.11. F-test for Equality of Several Means.** This test is carried out by the technique of Analysis of Variance, which plays a very important and fundamental role in Design of Experiments in Agricultural Statistics.

**14.5. Non-Central F-distribution.** The ratio of two independent  $\chi^2$  variates each divided by the corresponding d.f. has a non-central  $F$ -distribution if the numerator has a non-central  $\chi^2$ -distribution and the denominator has a central  $\chi^2$ -distribution. Thus, if  $X$  has a non-central  $\chi^2$ -distribution with  $n_1$  d.f. and non-centrality parameter  $\lambda$ , i.e., if  $X \sim \chi^2_{n_1}$  and  $Y$  is an independent (central)  $\chi^2$ -variate with  $n_2$  d.f. i.e., if  $Y \sim \chi^2_{n_2}$ , then the non-central  $F$ -statistic is defined as :

$$F' = \frac{X/n_1}{Y/n_2} = \frac{n_2 X}{n_1 Y}$$

**p.d.f. of  $F'$ .** Since  $X$  and  $Y$  are independent, their joint p.d.f. is given by .

$$p(x, y) = p_1(x) \cdot p_2(y) = \left[ \sum_{i=0}^{\infty} \frac{e^{-\lambda} \lambda^i}{i!} \cdot \frac{e^{-x/2} x^{(n_1/2)+i-1}}{2^{(n_1+2i)/2} \Gamma[(n_1+2i)/2]} \right] \times \frac{e^{-y/2} y^{(n_2/2)-1}}{2^{n_2/2} \Gamma(n_2/2)} ; \quad 0 \leq (x, y) < \infty.$$

Let us transform to the new set of r.v.'s  $F'$  and  $U$  defined by the transformation ;

$$F' = \frac{n_2 x}{n_1 y} \quad \text{and} \quad u = y \quad \Rightarrow \quad y = u, \quad x = \frac{n_1 u F'}{n_2}$$

$$J = \frac{\partial(x, y)}{\partial(F', u)} = \begin{vmatrix} \frac{n_1}{n_2} u & \frac{n_1}{n_2} F' \\ 0 & 1 \end{vmatrix} = \frac{n_1}{n_2} u$$

The joint p.d.f. of  $F'$  and  $U$  is given by

$$\begin{aligned} g(F', u) &= \left\{ \sum_{i=0}^{\infty} \frac{e^{-\lambda} \lambda^i}{i!} \frac{\exp \left[ -\frac{n_1 u F'}{2 n_2} \right] \cdot \left( \frac{n_1 u F'}{n_2} \right)^{(n_1/2)+i-1}}{2^{i+(n_1+n_2)/2} \Gamma(n_2/2) \Gamma[(n_1+2i)/2]} \right\} \\ &\quad \times e^{-u/2} u^{(n_2/2)-1} \cdot \left( \frac{n_1}{n_2} \right) u \\ &= \frac{n_1}{n_2} \sum_{i=0}^{\infty} \left\{ \frac{e^{-\lambda} \lambda^i}{i!} \frac{\left( \frac{n_1}{n_2} F' \right)^{(n_1/2)+i-1}}{2^{i+(n_1+n_2)/2} \Gamma(n_2/2) \Gamma[(n_1+2i)/2]} \right. \\ &\quad \times \left. \exp \left[ -\frac{u}{2} \left( 1 + \frac{n_1}{n_2} F' \right) \right] \cdot u^{\frac{n_1+n_2}{2}+i-1} \right\} \\ &\quad 0 \leq F' < \infty, 0 < u < \infty \end{aligned}$$

Integrating it w.r.t.  $u$  between the limits 0 to  $\infty$  and using Gamma Integral, we obtain the marginal p.d.f. of  $F'$  as

$$\begin{aligned}
 g(F') &= \frac{n_1}{n_2} \sum_{i=0}^{\infty} \left\{ \frac{e^{-\lambda} \lambda^i}{i!} \cdot \frac{\left(\frac{n_1}{n_2} F'\right)^{(n_1/2) + i - 1}}{2^{i + (n_1 + n_2)/2} \Gamma(n_2/2) \Gamma[(n_1 + 2i)/2]} \right. \\
 &\quad \times \left. \frac{\Gamma\left(\frac{n_1 + n_2}{2} + i\right)}{\left[\frac{1}{2} \left(1 + \frac{n_1}{n_2} F'\right)\right]^{i + (n_1 + n_2)/2}} \right\} \\
 &= \frac{n_1}{n_2} \sum_{i=0}^{\infty} \left\{ \frac{e^{-\lambda} \lambda^i}{i!} \cdot \frac{\left(\frac{n_1}{n_2} F'\right)^{(n_1/2) + i - 1}}{B\left(\frac{n_1}{2} + i, \frac{n_2}{2}\right)} \right. \\
 &\quad \times \left. \frac{1}{\left(1 + \frac{n_1}{n_2} F'\right)^{i + (n_1 + n_2)/2}} \right\}; \quad 0 \leq F' < \infty \quad \dots(14.23)
 \end{aligned}$$

**Remarks.** 1. For  $\lambda = 0$ , we get

$$g(F') = \frac{n_1}{n_2} \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{\left(\frac{n_1}{n_2} F'\right)^{(n_1/2) - 1}}{\left(1 + \frac{n_1}{n_2} F'\right)^{(n_1 + n_2)/2}}; \quad 0 \leq F' < \infty,$$

since for  $\lambda = 0$ , we get the contribution from the sum only when  $i = 0$  and all other terms vanish. Thus for  $\lambda = 0$ ,  $g(F')$  reduces to the p.d.f. of central  $F$ -distribution with  $(n_1, n_2)$  d.f.

2. The hyper-geometric function of first kind is defined by

$${}_1F_1(a, b, y) = \sum_{i=0}^{\infty} \frac{\Gamma(a+i) \Gamma b}{\Gamma a \Gamma(b+i)} \cdot \frac{y^i}{i!} \quad \dots(14.23b)$$

$$\begin{aligned}
 \therefore {}_1F_1\left(\frac{n_1 + n_2}{2}, \frac{n_1}{2}, \frac{\lambda n_1 F'}{n_2 \left(1 + \frac{n_1}{n_2} F'\right)}\right) \\
 &= \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{n_1 + n_2}{2} + i\right) \Gamma\left(\frac{n_1}{2}\right)}{\Gamma\left(\frac{n_1 + n_2}{2}\right) \Gamma\left(\frac{n_1}{2} + i\right)} \times \frac{(\lambda n_1 F')^i}{\left[n_2 \left(1 + \frac{n_1}{n_2} F'\right)\right]^i} \times \frac{1}{i!}
 \end{aligned}$$

$$\begin{aligned}
 &= \sum_{i=0}^{\infty} \frac{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)}{B\left(\frac{n_1}{2} + i, \frac{n_2}{2}\right)} \cdot \frac{\lambda^i}{i!} \frac{\left(\frac{n_1}{n_2}\right)^i F'^i}{\left(1 + \frac{n_1}{n_2} F'\right)^i} \\
 &\quad e^{-\lambda} \cdot \left(\frac{n_1}{n_2}\right)^{n_1/2} (F')^{\frac{n_1}{2}-1} \\
 \therefore g(F') = & \frac{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right) \cdot \left(1 + \frac{n_1}{n_2} F'\right)^{\frac{n_1+n_2}{2}}}{\times {}_1F_1\left(\frac{n_1+n_2}{2}, \frac{n_1}{2}, \frac{\lambda n_1 F'}{n_2 \left(1 + \frac{n_1}{n_2} F'\right)}\right)}
 \end{aligned}$$

3 It can be easily proved that the mean of  $F'_{n_1, n_2}$  is given by

$$\begin{aligned}
 E(F') &= \int_0^\infty F' g(F') dF' \\
 &= \sum_{i=0}^{\infty} \left[ \frac{e^{-\lambda} \lambda^i}{i!} \cdot \frac{n_2 (n_1 + 2i)}{n_1 (n_2 - 2)} \right]; n_2 > 2. \quad \dots(14-23c)
 \end{aligned}$$

If  $\lambda = 0$  (in which case we get contribution from the sum only when  $i = 0$ ), we get  $E(F') = \frac{n_2}{n_2 - 2}$ ,  $\dots(14-23d)$

which is the mean of central  $F$ -distribution with  $(n_1, n_2)$  d.f.

**14-7. Fisher's z-distribution.** In G.W. Snedecor's  $F$ -distribution with  $(v_1, v_2)$  d.f., if we put

$$F = \exp(2Z) \Rightarrow Z = \frac{1}{2} \log_e F \quad \dots(14-24)$$

The distribution of  $Z$  becomes:

$$\begin{aligned}
 g(z) &= p(F) \cdot \left| \frac{dF}{dz} \right| \\
 &= \frac{(v_1/v_2)^{v_1/2}}{B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \cdot \frac{(e^{2z})^{(v_1/2)} - 1}{\left[1 + \frac{v_1}{v_2} e^{2z}\right]^{(v_1+v_2)/2}} \cdot 2e^{2z} \\
 &= 2 \frac{(v_1/v_2)^{v_1/2}}{B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \cdot \frac{e^{v_1 z}}{\left[1 + \frac{v_1}{v_2} e^{2z}\right]^{(v_1+v_2)/2}}; -\infty < z < \infty \quad \dots(14-25)
 \end{aligned}$$

which is the probability function of Fisher's  $z$ -distribution with  $(v_1, v_2)$  d.f. The tables of significant values  $z_0$  of  $z$  which will be exceeded in random sampling with probabilities 0.05 and 0.01, i.e.,  $P(z > z_0) = 0.05$  and  $P(z > z_0') = 0.01$

corresponding to various d.f. ( $v_1, v_2$ ) were published by Fisher (c.f. Statistical Methods for Research Workers) in 1925. From these tables, Snedecor (1934-38) by using (14.24) deduced the tables of significant values of the variance ratio which he denoted by  $F$  in honour of Prof. R.A. Fisher.

**Remark.** With the help of relation (14.24), all the applications of  $F$ -distribution may be regarded as the applications of  $z$ -distribution also.

### 14.7-1. Moment Generating Function of $z$ -distribution.

$$\begin{aligned} M_Z(t) &= E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} g(z) dz \\ &= \int_0^{\infty} F^{v_2/2} f(F) dF \quad [\because e^{2x} = F] \end{aligned}$$

Since  $\mu_r'$  (about origin) for  $F$ -distribution is  $\int_0^{\infty} F^r f(F) dF$ , we can find

m.g.f. of the  $z$ -distribution by putting  $r = t/2$  in the expression for  $\mu_r'$  for  $F$ -distribution.

$$\begin{aligned} \text{Hence } M_Z(t) &= \left( \frac{v_2}{v_1} \right)^{v_2/2} \cdot \frac{\Gamma((v_1 + t)/2)}{\Gamma(v_1/2)} \frac{\Gamma((v_2 - t)/2)}{\Gamma(v_2/2)} \quad [\text{c.f. Equation (14.15)}] \\ \Rightarrow \quad K_Z(t) &= \log M_Z(t) \\ &= \frac{t}{2} [\log v_2 - \log v_1] + \log \Gamma((v_1 + t)/2) \\ &\quad + \log \Gamma((v_2 - t)/2) - \log \Gamma(v_1/2) - \log \Gamma(v_2/2) \end{aligned}$$

Using Stirling's approximation for  $n!$ , when  $n$  is large, viz.,

$$\begin{aligned} \lim_{n \rightarrow \infty} \Gamma(n+1) &= \lim_{n \rightarrow \infty} n! \approx \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}} \\ \Rightarrow \quad \log \Gamma(n+1) &= (n + \frac{1}{2}) \log n - n + \log \sqrt{2\pi}, \text{ we get} \\ \kappa_1 = \mu_1' &= \frac{1}{2} \left( \frac{1}{v_2} - \frac{1}{v_1} \right) \\ \kappa_2 = \mu_2 &= \frac{1}{2} \left( \frac{1}{v_2} + \frac{1}{v_1} + \frac{1}{v_1^2} + \frac{1}{v_2^2} \right) \\ \kappa_3 = \mu_3 &= \frac{1}{2} \left[ \left( \frac{1}{v_2^2} - \frac{1}{v_1^2} \right) + \left( \frac{1}{v_2^3} + \frac{1}{v_1^3} \right) \right] \\ \kappa_4 = \mu_4 - 3\mu_2^2 &= \frac{1}{v_1^3} + \frac{1}{v_2^3} + 3 \left( \frac{1}{v_1^4} + \frac{1}{v_2^4} \right), \end{aligned}$$

whence  $\beta_1$  and  $\beta_2$  can be found.

**Remark.** *z*-distribution tends to normal distribution with mean  $\frac{1}{2} \left( \frac{1}{v_2} - \frac{1}{v_1} \right)$  and variance  $\frac{1}{2} \left( \frac{1}{v_1} + \frac{1}{v_2} \right)$ , as  $v_1$  and  $v_2$  become large.

**14.8. Fisher's *z*-transformation.** To test the significance of an observed sample correlation coefficient from an uncorrelated bivariate normal population, *t*-test (*cf.* § 14.2-10) is used. But in random sample of size  $n$ , from a bivariate normal population in which  $\rho \neq 0$ , Prof. R.A. Fisher proved that the distribution of '*r*' is by no means normal and in the neighbourhood of  $\rho = \pm 1$ , its probability curve is extremely skewed even for large  $n$ . If  $\rho \neq 0$ , Fisher suggested the following transformation

$$Z = \frac{1}{2} \log_e \frac{1+r}{1-r} = \tanh^{-1} r \quad \dots(14.26)$$

and proved that even for small samples, the distribution of *Z* is approximately normal with mean

$$\xi = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho} = \tanh^{-1} \rho \quad \dots[14.26(a)]$$

and variance  $1/(n-3)$  and for large values of  $n$ , say  $> 50$ , the approximation is fairly good.

*z*-transformation has the following applications in Statistics.

(1) To test if an observed value of '*r*' differs significantly from a hypothetical value  $\rho$  of the population correlation coefficient.

$H_0$ : There is no significant difference between *r* and  $\rho$ . In other words, the given sample has been drawn from a bivariate normal population with correlation coefficient  $\rho$ .

If we take

$$Z = \frac{1}{2} \log_e \{(1+r)/(1-r)\} \text{ and } \xi = \frac{1}{2} \log_e \{(1+\rho)/(1-\rho)\},$$

then under  $H_0$ ,

$$Z \sim N\left(\xi, \frac{1}{n-3}\right) \Rightarrow \frac{Z - \xi}{\sqrt{1/(n-3)}} \sim N(0, 1)$$

Thus if  $(Z - \xi) \sqrt{(n-3)} > 1.96$ ,  $H_0$  is rejected at 5% level of significance and if it is greater than 2.58,  $H_0$  is rejected at 1% level of significance, where *Z* and  $\xi$  are defined in (14.26) and (14.26a).

**Remark.** *Z* defined in equation (14.26) should not be confused with the *Z* used in Fisher's *z*-distribution (*cf.* § 14.7).

**Example 14.24.** A correlation coefficient of 0.72 is obtained from a sample of 29 pairs of observations.

(i) Can the sample be regarded as drawn from a bivariate normal population in which true correlation coefficient is 0.8?

(ii) Obtain 95% confidence limits for  $\rho$  in the light of the information provided by the sample.

**Solution.** (i)  $H_0$ : There is no significant difference between  $r = 0.72$ ; and

$\rho = 0.80$ , i.e., the sample can be regarded as drawn from the bivariate normal population with  $\rho = 0.8$ .

Here  $Z = \frac{1}{2} \log_e \left( \frac{1+r}{1-r} \right) = 1.1513 \log_{10} \left( \frac{1+r}{1-r} \right)$   
 $= 1.1513 \log_{10} 6.14 = 0.907$

$$\xi = \frac{1}{2} \log_e \left( \frac{1+\rho}{1-\rho} \right) = 1.1513 \log_{10} \frac{(1+0.8)}{(1-0.8)}$$
  
 $= 1.1513 \times 0.9541 = 1.1$

$$\text{S.E.}(Z) = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{26}} = 0.196$$

Under  $H_0$ , the test statistic is

$$U = \frac{Z - \xi}{1/\sqrt{n-3}} \sim N(0, 1)$$

Now  $U = \frac{(0.907 - 1.100)}{0.196} = -0.985$

Since  $|U| < 1.96$ , it is not significant at 5% level of significance and  $H_0$  may be accepted. Hence the sample may be regarded as coming from a bivariate normal population with  $\rho = 0.8$ .

(ii) 95% confidence limits for  $\rho$  on the basis of the information supplied by the sample, are given by

$$|U| \leq 1.96$$

$$|Z - \xi| \leq 1.96 \times \frac{1}{\sqrt{n-3}} = 1.96 \times 0.196$$

$$\Rightarrow |0.907 - \xi| \leq 0.384$$

$$\Rightarrow 0.907 - 0.384 \leq \xi \leq 0.907 + 0.384$$

$$\Rightarrow 0.523 \leq \xi \leq 1.291$$

$$\Rightarrow 0.523 \leq \frac{1}{2} \log_e \left( \frac{1+\rho}{1-\rho} \right) \leq 1.291$$

$$\Rightarrow 0.523 \leq 0.1513 \log_{10} \left( \frac{1+\rho}{1-\rho} \right) \leq 1.291$$

$$\Rightarrow \frac{0.523}{0.1513} \leq \log_{10} \left( \frac{1+\rho}{1-\rho} \right) \leq \frac{1.291}{0.1513}$$

$$\Rightarrow 0.4543 \leq \log_{10} \left( \frac{1+\rho}{1-\rho} \right) \leq 1.1213 \quad \dots(*)$$

Now  $\log_{10} \left( \frac{1+\rho}{1-\rho} \right) = 0.4543$  and  $\log_{10} \left( \frac{1+\rho}{1-\rho} \right) = 1.1213$

$$\Rightarrow \frac{1+\rho}{1-\rho} = \text{Antilog}(0.4543) = 2.846 \quad \Rightarrow \frac{1+\rho}{1-\rho} = \text{Antilog}(1.1213) = 13.22$$

$$\Rightarrow \rho = \frac{2.846 - 1}{2.846 + 1} = \frac{1.846}{3.846} = 0.4799 \quad \Rightarrow \rho = \frac{13.22 - 1}{13.22 + 1} = \frac{12.22}{14.22} = 0.86$$

Hence, substituting in (\*) we get  $0.48 \leq \rho \leq 0.86$

(2) To test the significance of the difference between two independent sample correlation coefficients. Let  $r_1$  and  $r_2$  be the sample correlation coefficients observed in two independent samples of sizes  $n_1$  and  $n_2$  respectively then

$$Z_1 = \frac{1}{2} \log_e \left( \frac{1 + r_1}{1 - r_1} \right) \text{ and } Z_2 = \frac{1}{2} \log_e \left( \frac{1 + r_2}{1 - r_2} \right)$$

Under the null hypothesis  $H_0$ : that sample correlation coefficients do not differ significantly, i.e., the samples are drawn from the same bivariate normal population or from different populations with same correlation coefficient  $\rho$ , (say), the statistic

$$Z = \frac{(Z_1 - Z_2) - E(Z_1 - Z_2)}{\text{S.E.}(Z_1 - Z_2)} \sim N(0, 1)$$

$$\text{Now } E(Z_1 - Z_2) = E(Z_1) - E(Z_2) = \xi_1 - \xi_2 = 0$$

$$\left[ \because \xi_1 = \xi_2 = \frac{1}{2} \log_e \frac{1 + \rho}{1 - \rho} \text{ (under } H_0) \right]$$

$$\text{and } \text{S.E.}(Z_1 - Z_2) = \sqrt{V(Z_1) + V(Z_2)}$$

[Covariance term vanishes since samples are independent]

$$= \sqrt{\left\{ \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \right\}}$$

Under  $H_0$ , the test statistic is

$$Z = \frac{Z_1 - Z_2}{\sqrt{\left\{ \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \right\}}} \sim N(0, 1)$$

By comparing this value with 1.96 or 2.58,  $H_0$  may be accepted or rejected at 5% and 1% level of significance respectively.

(3) To obtain pooled estimate of  $\rho$ . Let  $r_1, r_2, \dots, r_k$  be observed correlation coefficients in  $k$ -independent samples of sizes  $n_1, n_2, \dots, n_k$  from a bivariate normal population. The problem is to combine these estimates of  $\rho$  to get a pooled estimate for the parameter.

$$\text{If we take } Z_i = \frac{1}{2} \log_e \left( \frac{1 + r_i}{1 - r_i} \right); i = 1, 2, \dots, k$$

then  $Z_i; i = 1, 2, \dots, k$  are independent normal variates with variances  $\frac{1}{(n_i - 3)}$ ;  $i = 1, 2, \dots, k$  and common mean

$$\xi = \frac{1}{2} \log_e \left( \frac{1 + \rho}{1 - \rho} \right)$$

The weighted mean (say  $\bar{Z}$ ) of these  $Z_i$ 's is given by

$$\bar{Z} = \sum_{i=1}^k w_i Z_i / \sum_{i=1}^k w_i,$$

where  $w_i$  is the weight of  $Z_i$ .

Now  $\bar{Z}$  is also an unbiased estimate of  $\xi$ , since

$$E(\bar{Z}) = \frac{1}{\sum w_i} \left[ E \left( \sum_{i=1}^k w_i Z_i \right) \right] = \frac{1}{\sum w_i} \left[ \sum w_i E(Z_i) \right] = \frac{1}{\sum w_i} \left[ \sum w_i \xi \right] = \xi$$

$$\text{and } V(\bar{Z}) = \frac{1}{(\sum w_i)^2} V[\sum w_i Z_i] = \frac{1}{(\sum w_i)^2} [\sum w_i^2 V(Z_i)]$$

The weights  $w_i$ 's, ( $i = 1, 2, \dots, n$ ) are so chosen that  $\bar{Z}$  has minimum variance.

In order that  $V(\bar{Z})$  is minimum for variations in  $w_i$ , we should have

$$\begin{aligned} \frac{\partial}{\partial w_i} V(\bar{Z}) &= 0; \quad i = 1, 2, \dots, k \\ \Rightarrow \frac{(\sum w_i)^2 2w_i V(Z_i) - [\sum_i w_i^2 V(Z_i)] 2(\sum w_i)}{(\sum w_i)^4} &= 0 \\ \Rightarrow w_i V(Z_i) &= \frac{\sum w_i^2 V(Z_i)}{\sum w_i}, \text{ a constant.} \\ \therefore w_i &\propto \frac{1}{V(Z_i)} = (n_i - 3); \quad i = 1, 2, \dots, k \end{aligned} \quad \dots (*)$$

Hence the minimum variance estimate of  $\xi$  is given by

$$\bar{Z} = \frac{\sum_{i=1}^k w_i Z_i}{\sum_{i=1}^k w_i} = \frac{\sum_{i=1}^k (n_i - 3) Z_i}{\sum_{i=1}^k (n_i - 3)} \quad [\text{On using (*)}]$$

and the best estimate of  $\rho$  is then given by

$$\bar{Z} = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho} \Rightarrow \hat{\rho} = \tanh \left[ \frac{\sum (n_i - 3) Z_i}{\sum (n_i - 3)} \right] \quad (\text{c.f. } \S \text{ 13.9.1})$$

**Remark.** Minimum variance of  $\bar{Z}$  is given by

$$\{V(\bar{Z})\}_{\min} = \frac{\sum \left\{ (n_i - 3)^2 \left( \frac{1}{n_i - 3} \right) \right\}}{(\sum (n_i - 3))^2} = \frac{\sum (n_i - 3)}{(\sum (n_i - 3))^2} = \frac{1}{\sum_{i=1}^k (n_i - 3)}$$

# **Statistical Inference-I (Theory of Estimation)**

**15-1. Introduction.** The theory of estimation was founded by Prof. R.A. Fisher in a series of fundamental papers round about 1930.

**Parameter Space.** Let us consider a random variable  $X$  with p.d.f.  $f(x, \theta)$ . In most common applications, though not always, the functional form of the population distribution is assumed to be known except for the value of some unknown parameter(s)  $\theta$  which may take any value on a set  $\Theta$ . This is expressed by writing the p.d.f. in the form  $f(x, \theta), \theta \in \Theta$ . The set  $\Theta$ , which is the set of all possible values of  $\theta$  is called the *parameter space*. Such a situation gives rise not to one probability distribution but a family of probability distributions which we write as  $\{f(x, \theta), \theta \in \Theta\}$ . For example if  $X \sim N(\mu, \sigma^2)$ , then the parameter space

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty ; 0 < \sigma^2 < \infty\}$$

In particular, for  $\sigma^2 = 1$ , the family of probability distributions is given by

$$\{N(\mu, 1) ; \mu \in \Theta\}, \text{ where } \Theta = \{\mu : -\infty < \mu < \infty\}$$

In the following discussion we shall consider a general family of distributions

$$\{f(x_i ; \theta_1, \theta_2, \dots, \theta_k) : \theta_i \in \Theta, i = 1, 2, \dots, k\}.$$

Let us consider a random sample  $x_1, x_2, \dots, x_n$  of size  $n$  from a population, with probability function  $f(x ; \theta_1, \theta_2, \dots, \theta_k)$ , where  $\theta_1, \theta_2, \dots, \theta_k$  are the unknown population parameters. There will then always be an infinite number of functions of sample values, called statistics, which may be proposed as estimates of one or more of the parameters.

Evidently, the best estimate would be one that falls nearest to the true value of the parameter to be estimated. In other words, the statistic whose distribution concentrates as closely as possible near the true value of the parameter may be regarded the best estimate. Hence the basic problem of the estimation in the above case, can be formulated as follows :

'We wish to determine the functions of the sample observations :

$T_1 = \hat{\theta}_1(x_1, x_2, \dots, x_n), T_2 = \hat{\theta}_2(x_1, x_2, \dots, x_n), \dots, T_k = \hat{\theta}_k(x_1, x_2, \dots, x_n)$ , such that their distribution is concentrated as closely as possible near the true value of the parameter.

The estimating functions are then referred to as *estimators*.

**15-2. Characteristics of Estimators.** The following are some of the criteria that should be satisfied by a good estimator.

(i) *Consistency*

- (ii) *Unbiasedness*
- (iii) *Efficiency* and
- (iv) *Sufficiency*

We shall now, briefly, explain these terms one by one.

**15.3. Consistency.** An estimator  $T_n = T(x_1, x_2, \dots, x_n)$ , based on a random sample of size  $n$ , is said to be consistent estimator of  $\gamma(\theta)$ ,  $\theta \in \Theta$ , the parameter space, if  $T_n$  converges to  $\gamma(\theta)$  in probability.

i.e., if

$$T_n \xrightarrow{P} \gamma(\theta) \text{ as } n \rightarrow \infty \quad \dots(15.1)$$

In other words,  $T_n$  is a consistent estimator of  $\gamma(\theta)$  if for every  $\epsilon > 0$ ,  $\eta > 0$ , there exists a positive integer  $n \geq m(\epsilon, \eta)$  such that

$$P [|T_n - \gamma(\theta)| < \epsilon] \rightarrow 1 \text{ as } n \rightarrow \infty \quad \dots(15.2)$$

$$\Rightarrow P [|T_n - \gamma(\theta)| < \epsilon] > 1 - \eta ; \forall n \geq m \quad \dots(15.2a)$$

where  $m$  is some very large value of  $n$ .

**Remark.** If  $X_1, X_2, \dots, X_n$  is a random sample from a population with finite mean  $EX_i = \mu < \infty$ , then by Khinchine's weak law of large numbers (W.L.L.N), we have

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} E(X_i) = \mu, \text{ as } n \rightarrow \infty.$$

Hence sample mean ( $\bar{X}_n$ ) is always a consistent estimator of the population mean ( $\mu$ ).

**15.4. Unbiasedness.** Obviously, consistency is a property concerning the behaviour of an estimator for indefinitely large values of the sample size  $n$ , i.e., as  $n \rightarrow \infty$ . Nothing is regarded of its behaviour for finite  $n$ .

Moreover, if there exists a consistent estimator, say,  $T_n$  of  $\gamma(\theta)$ , then infinitely many such estimators can be constructed, e.g.,

$$T_n' = \left( \frac{n-a}{n-b} \right) T_n = \left[ \frac{1-(a/n)}{1-(b/n)} \right] T_n \xrightarrow{P} \gamma(\theta), \text{ as } n \rightarrow \infty$$

and hence, for different values of  $a$  and  $b$ ,  $T_n'$  is also consistent for  $\gamma(\theta)$ .

Unbiasedness is a property associated with finite  $n$ . A statistic

$T_n = T(x_1, x_2, \dots, x_n)$ , is said to be an unbiased estimator of  $\gamma(\theta)$  if

$$E(T_n) = \gamma(\theta), \text{ for all } \theta \in \Theta \quad \dots(15.3)$$

We have seen (c.f. § 12.12) that in sampling from a population with mean  $\mu$  and variance  $\sigma^2$ ,

$$E(\bar{x}) = \mu \text{ and } E(s^2) \neq \sigma^2 \text{ but } E(S^2) = \sigma^2.$$

Hence there is a reason to prefer

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \text{ to the sample variance } s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

**Remark.** If  $E(T_n) > \theta$ ,  $T_n$  is said to be positively biased and if  $E(T_n) < \theta$ , it is said to be negatively biased, the amount of bias  $b(\theta)$  being given by

$$b(\theta) = E(T_n) - \gamma(\theta), \theta \in \Theta \quad \dots(15.3a)$$

#### 15.4.1. Invariance Property of Consistent Estimators.

**Theorem 15.1.** If  $T_n$  is a consistent estimator of  $\gamma(\theta)$  and  $\psi(\gamma(\theta))$  is a continuous function of  $\gamma(\theta)$ , then  $\psi(T_n)$  is a consistent estimator of  $\psi(\gamma(\theta))$ .

**Proof.** Since  $T_n$  is a consistent estimator of  $\gamma(\theta)$ ,  $T_n \xrightarrow{P} \gamma(\theta)$  as  $n \rightarrow \infty$  i.e., for every  $\epsilon > 0$ ,  $\eta > 0$ ,  $\exists$  a positive integer  $m \geq m(\epsilon, \eta)$  such that

$$P[|T_n - \gamma(\theta)| < \epsilon] > 1 - \eta, \forall n \geq m \quad \dots(*)$$

Since  $\psi(\cdot)$  is a continuous function, for every  $\epsilon > 0$ , however small,  $\exists$  a positive number  $\epsilon_1$  such that  $|\psi(T_n) - \psi(\gamma(\theta))| < \epsilon_1$ , whenever  $|T_n - \gamma(\theta)| < \epsilon$

$$\text{i.e., } |T_n - \gamma(\theta)| < \epsilon \Rightarrow |\psi(T_n) - \psi(\gamma(\theta))| < \epsilon_1 \quad \dots(**)$$

For two events  $A$  and  $B$ ,

$$\text{if } A \Rightarrow B, \text{ then } A \subseteq B \Rightarrow P(A) \leq P(B) \Rightarrow P(B) \geq P(A) \quad \dots(***)$$

From (\*\*) and (\*\*\*) we get

$$P[|\psi(T_n) - \psi(\gamma(\theta))| < \epsilon_1] \geq P[|T_n - \gamma(\theta)| < \epsilon]$$

$$\Rightarrow P[|\psi(T_n) - \psi(\gamma(\theta))| < \epsilon_1] \geq 1 - \eta ; \forall n \geq m \quad [\text{Using } (*)]$$

$$\Rightarrow \psi(T_n) \xrightarrow{P} \psi(\gamma(\theta)), \text{ as } n \rightarrow \infty$$

$\psi(T_n)$  is a consistent estimator of  $\psi(\gamma(\theta))$ .

#### 15.4.2. Sufficient Conditions for Consistency.

**Theorem 15.2.** Let  $\{T_n\}$  be a sequence of estimators such that for all  $\theta \in \Theta$ ,

$$(i) E_\theta(T_n) \rightarrow \gamma(\theta), n \rightarrow \infty$$

$$\text{and} \quad (ii) \text{Var}_\theta(T_n) \rightarrow 0, \text{ as } n \rightarrow \infty.$$

Then  $T_n$  is a consistent estimator of  $\gamma(\theta)$ .

**Proof.** We have to prove that  $T_n$  is a consistent estimator of  $\gamma(\theta)$

$$\text{i.e., } T_n \xrightarrow{P} \gamma(\theta), \text{ as } n \rightarrow \infty$$

$$\text{i.e., } P[|T_n - \gamma(\theta)| < \epsilon] > 1 - \eta ; \forall n \geq m (\epsilon, \eta) \quad \dots(15.4)$$

where  $\epsilon$  and  $\eta$  are arbitrarily small positive numbers and  $m$  is some large value of  $n$ :

Applying Chebychev's inequality to the statistic  $T_n$ , we get

$$P[|T_n - E_\theta(T_n)| \leq \delta] \geq 1 - \frac{\text{Var}_\theta(T_n)}{\delta^2} \quad \dots(15.5)$$

We have

$$|T_n - \gamma(\theta)| = |T_n - E(T_n) + E(T_n) - \gamma(\theta)|$$

$$\leq |T_n - E_\theta(T_n)| + |E_\theta(T_n) - \gamma(\theta)| \quad \dots(15.6)$$

Now

$$|T_n - E_\theta(T_n)| \leq \delta \Rightarrow |T_n - \gamma(\theta)| \leq \delta + |E_\theta(T_n) - \gamma(\theta)| \quad \dots(15.7)$$

Hence, on using (\*\*\*) of Theorem 15.1, we get

$$\begin{aligned} P[|T_n - \gamma(\theta)| \leq \delta + |E_\theta(T_n) - \gamma(\theta)|] &\geq P[|T_n - E_\theta(T_n)| \leq \delta] \\ &\geq 1 - \frac{\text{Var}_\theta(T_n)}{\delta^2} \quad [\text{From (15.5)}] \dots(15.8) \end{aligned}$$

We are given :

$$E_\theta(T_n) \rightarrow \gamma(\theta) \quad \forall \theta \in \Theta \text{ as } n \rightarrow \infty.$$

Hence, for every  $\delta_1 > 0$ ,  $\exists$  a positive integer  $n \geq n_0(\delta_1)$  such that

$$|E_\theta(T_n) - \gamma(\theta)| \leq \delta_1, \quad \forall n \geq n_0(\delta_1) \quad \dots(15.9)$$

Also  $\text{Var}_\theta(T_n) \rightarrow 0$  as  $n \rightarrow \infty$ . (Given).

$$\therefore \frac{\text{Var}_\theta(T_n)}{\delta^2} \leq \eta, \quad \forall n \geq n_0'(\eta) \quad \dots(15.10)$$

where  $\eta$  is arbitrarily small positive number.

Substituting from (15.9) and (15.10) in (15.8), we get

$$P[|T_n - \gamma(\theta)| \leq \delta + \delta_1] \geq 1 - \eta; \quad n \geq m(\delta_1, \eta)$$

$$\Rightarrow P[|T_n - \gamma(\theta)| \leq \varepsilon] \geq 1 - \eta; \quad n \geq m$$

where  $m = \max(n_0, n_0')$  and  $\varepsilon = \delta + \delta_1 > 0$ .

$$\Rightarrow T_n \xrightarrow{P} \gamma(\theta), \quad \text{as } n \rightarrow \infty \quad [\text{Using (15.4)}]$$

$\Rightarrow T_n$  is a consistent estimator of  $\gamma(\theta)$ .

**Example 15.1.**  $x_1, x_2, \dots, x_n$  is a random sample from a normal population  $N(\mu, 1)$ . Show that  $t = \frac{1}{n} \sum_{i=1}^n x_i^2$ , is an unbiased estimator of  $\mu^2 + 1$ .

**Solution.** (a) We are given

$$E(x_i) = \mu, \quad V(x_i) = 1 \quad \forall i = 1, 2, \dots, n$$

$$\text{Now} \quad E(x_i^2) = V(x_i) + [E(x_i)]^2 = 1 + \mu^2$$

$$E(t) = E\left[\frac{1}{n} \sum_{i=1}^n x_i^2\right] = \frac{1}{n} \sum_{i=1}^n E(x_i^2) = \frac{1}{n} \sum_{i=1}^n (1 + \mu^2) = 1 + \mu^2$$

Hence  $t$  is an unbiased estimator of  $1 + \mu^2$ .

**Example 15.2.** If  $T$  is an unbiased estimator for  $\theta$ , show that  $T^2$  is a biased estimator for  $\theta^2$ .

**Solution.** Since  $T$  is an unbiased estimator for  $\theta$ , we have

$$E(T) = \theta$$

$$\text{Also} \quad \text{Var}(T) = E(T^2) - [E(T)]^2 = E(T^2) - \theta^2$$

$$\Rightarrow E(T^2) = \theta^2 + \text{Var}(T), (\text{Var } T > 0).$$

Since  $E(T^2) \neq \theta^2$ ,  $T^2$  is a biased estimator for  $\theta^2$ .

**Example 15.3.** Show that  $\frac{\sum x_i (\sum x_i - 1)}{n(n-1)}$  is an unbiased estimate of  $\theta^2$ , for the sample  $x_1, x_2, \dots, x_n$  drawn on  $X$  which takes the values 1 or 0 with respective probabilities  $\theta$  and  $(1-\theta)$ .

**Solution.** Since ' $x_1, x_2, \dots, x_n$ ' is a random sample from Bernoulli population with parameter  $\theta$ ,

$$T = \sum_{i=1}^n x_i \sim B(n, \theta)$$

$$\Rightarrow E(T) = n\theta \text{ and } \text{Var}(T) = n\theta(1-\theta)$$

$$\begin{aligned} E\left[\frac{\sum x_i (\sum x_i - 1)}{n(n-1)}\right] &= E\left[\frac{T(T-1)}{n(n-1)}\right] \\ &= \frac{1}{n(n-1)} [E(T^2) - E(T)] \\ &= \frac{1}{n(n-1)} [\text{Var}(T) + \{E(T)\}^2 - E(T)] \\ &= \frac{1}{n(n-1)} [n\theta(1-\theta) + n^2\theta^2 - n\theta] \\ &= \frac{n\theta^2(n-1)}{n(n-1)} = \theta^2 \end{aligned}$$

$\Rightarrow [\sum x_i (\sum x_i - 1)] / [n(n-1)]$  is an unbiased estimator of  $\theta^2$ .

**Example 15.4.** Let  $X$  be distributed in the Poisson form with parameter  $\theta$ . Show that the only unbiased estimator of  $\exp[-(k+1)\theta]$ ,  $k > 0$ , is  $T(X) = (-k)^X$  so that

$T(x) > 0$  if  $x$  is even

and  $T(x) < 0$  if  $x$  is odd.

[Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1988]

$$\begin{aligned} \text{Solution. } E\{T(X)\} &= E[(-k)^X], k > 0 = \sum_{x=0}^{\infty} (-k)^x \left\{ \frac{e^{-\theta}}{x!} \theta^x \right\} \\ &= e^{-\theta} \sum_{x=0}^{\infty} \left[ \frac{(-k\theta)^x}{x!} \right] = e^{-\theta} \cdot e^{-k\theta} = e^{-(1+k)\theta} \end{aligned}$$

$\Rightarrow T(X) = (-k)^X$  is an unbiased estimator for  $\exp[-(1+k)\theta]$ ,  $k > 0$ .

**Example 15.5.** (a) Prove that in sampling from a  $N(\mu, \sigma^2)$  population, the sample mean is a consistent estimator of  $\mu$ .

(b) Prove that for Cauchy's distribution not sample mean but sample median is a consistent estimator of the population mean.

**Solution.** In sampling from a  $N(\mu, \sigma^2)$  population, the sample mean  $\bar{x}$  is also normally distributed as  $N(\mu, \sigma^2/n)$ .

$$\Rightarrow E(\bar{x}) = \mu \text{ and } V(\bar{x}) = \sigma^2/n$$

Thus as  $n \rightarrow \infty$ ,

$$E(\bar{x}) = \mu \text{ and } V(\bar{x}) = 0$$

Hence by Theorem 15.2,  $\bar{x}$  is a consistent estimator for  $\mu$ .

(b) The Cauchy's population is given by the probability function

$$dF(x) = \frac{1}{\pi} \cdot \frac{dx}{1 + (x - \mu)^2}, -\infty < x < \infty$$

The mean of the distribution, if we conventionally agree to assume that it exists, is at  $x = \mu$ .

If  $\bar{x}$ , the sample mean is taken as an estimator of  $\mu$ , then the sampling distribution of  $\bar{x}$  is given by

$$dF(\bar{x}) = \frac{1}{\pi} \cdot \frac{d\bar{x}}{1 + (\bar{x} - \mu)^2}; -\infty < \bar{x} < \infty \quad \dots(i)$$

because in Cauchy's distribution, the distribution of  $\bar{x}$  is same as the distribution of  $x$ .

Since in this case, the distribution of  $\bar{x}$  is same as the distribution of any single sample observation, it does not increase in accuracy with increasing  $n$ . Hence we have

$$E(\bar{x}) = \mu \text{ but } V(\bar{x}) = V(x) \neq 0, \text{ as } n \rightarrow \infty$$

Hence by Theorem 15.2,  $\bar{x}$  is not a consistent estimator of  $\mu$  in this case.

Consideration of symmetry of (i) is enough to show that the sample median  $Md$  is an unbiased estimate of the population mean, which of course is same as the population median.

$$\therefore E(Md) = \mu \quad \dots(ii)$$

For large  $n$ , the sampling distribution of median is asymptotically normal and is given by

$$dF \propto \exp [-2n f_1^2 (x - \mu)^2] dx$$

where  $f_1$  is the median ordinate of the parent population.

$$\text{i.e., } dF \propto \exp \left\{ -\frac{(x - \mu)^2}{1/(2n f_1^2)} \right\} \quad \dots(iii)$$

But  $f_1 = \text{Median ordinate of (i)}$

$= \text{Modal ordinate of (i)}$  [Because of symmetry]

$$= [f(x)]_{x=\mu} = \frac{1}{\pi}$$

Hence, from (iii), the variance of the sampling distribution of median is :

$$V(Md) = \frac{1}{4n f_1^2} = \frac{1}{4n(1/\pi)^2} = \frac{\pi^2}{4n} \rightarrow 0 \text{ as } n \rightarrow \infty \quad \dots(iv)$$

Hence from (ii) and (iv), using Theorem 15-2, we conclude that for Cauchy's distribution, median is a consistent estimator for  $\mu$ .

**Example 15-6.** If  $X_1, X_2, \dots, X_n$  are random observations on a Bernoulli variate  $X$  taking the value 1 with probability  $p$  and the value 0 with probability  $(1-p)$ , show that :

$\frac{\sum x_i}{n} \left(1 - \frac{\sum x_i}{n}\right)$  is a consistent estimator of  $p(1-p)$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

**Solution.** Since  $X_1, X_2, \dots, X_n$  are i.i.d Bernoulli variates with parameter ' $p$ ',

$$T = \sum_{i=1}^n x_i \sim B(n, p)$$

$$\Rightarrow E(T) = np \quad \text{and} \quad \text{Var}(T) = npq$$

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

$$\therefore E(\bar{X}) = \frac{1}{n} E(T) = \frac{1}{n} \cdot np = p$$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{T}{n}\right) = \frac{1}{n^2} \cdot \text{Var}(T) = \frac{pq}{n} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Since  $E(\bar{X}) \rightarrow p$  and  $\text{Var}(\bar{X}) \rightarrow 0$ , as  $n \rightarrow \infty$ ;  $\bar{X}$  is a consistent estimator of  $p$ .

Also  $\frac{\sum x_i}{n} \left(1 - \frac{\sum x_i}{n}\right) = \bar{X} (1 - \bar{X})$ , being a polynomial in  $\bar{X}$ , is a continuous function of  $\bar{X}$ .

Since  $\bar{X}$  is consistent estimator of  $p$ , by the invariance property of consistent estimators (Theorem 15-1),  $\bar{X} (1 - \bar{X})$  is a consistent estimator of  $p(1-p)$ .

**15-5. Efficient Estimators. Efficiency.** Even if we confine ourselves to unbiased estimates, there will, in general, exist more than one consistent estimator of a parameter. For example, in sampling from a normal population  $N(\mu, \sigma^2)$ , when  $\sigma^2$  is known, sample mean  $\bar{x}$  is an unbiased and consistent estimator of  $\mu$  [c.f. Example 15-5(a)].

From symmetry it follows immediately that sample median ( $Md$ ) is an unbiased estimate of  $\mu$ , which is the same as the population median. Also for large  $n$ ,

$$V(Md) = \frac{1}{4n f_1^2} \quad [c.f. \text{ Example 15-5(b)}]$$

Here

$$\begin{aligned}f_1 &= \text{Median ordinate of the parent distribution.} \\&= \text{Modal ordinate of the parent distribution.} \\&= \left[ \frac{1}{\sigma\sqrt{2\pi}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right) \right]_{x=\mu} = \frac{1}{\sigma\sqrt{2\pi}}\end{aligned}$$

$$\therefore V(Md) = \frac{1}{4n} \cdot 2\pi\sigma^2 = \frac{\pi\sigma^2}{2n}$$

Since  
and

$$\left. \begin{aligned}E(Md) &= \mu \\V(Md) &\rightarrow 0\end{aligned}\right\}, \text{ as } n \rightarrow \infty$$

median is also an unbiased and consistent estimator of  $\mu$ .

Thus, there is a necessity of some further criterion which will enable us to choose between the estimators with the common property of consistency. Such a criterion which is based on the variances of the sampling distribution of estimators is usually known as *efficiency*.

If, of the two consistent estimators  $T_1, T_2$  of a certain parameter  $\theta$ , we have

$$V(T_1) < V(T_2), \text{ for all } n \quad \dots(15.11)$$

then  $T_1$  is more efficient than  $T_2$  for all sample sizes.

We have seen above :

$$\text{For all } n, \quad V(\bar{x}) = \frac{\sigma^2}{n}$$

$$\text{and for large } n, \quad V(Md) = \frac{\pi\sigma^2}{2n} = 1.57 \frac{\sigma^2}{n}$$

Since  $V(\bar{x}) < V(Md)$ , we conclude that for normal distribution, sample mean is more efficient estimator for  $\mu$  than the sample median, for large samples at least.

**15.5.1. Most Efficient Estimator.** If in a class of consistent estimators for a parameter, there exists one whose sampling variance is less than that of any such estimator, it is called the most efficient estimator. Whenever such an estimator exists, it provides a criterion for measurement of efficiency of the other estimators.

**Efficiency (Def.)** If  $T_1$  is the most efficient estimator with variance  $V_1$  and  $T_2$  is any other estimator with variance  $V_2$ , then the efficiency  $E$  of  $T_2$  is defined as :

$$E = \frac{V_1}{V_2} \quad \dots(15.12)$$

Obviously,  $E$  cannot exceed unity.

If  $T, T_1, T_2, \dots, T_n$  are all estimators of  $\gamma(\theta)$  and  $\text{Var}(T)$  is minimum, then the efficiency  $E_i$  of  $T_i$ , ( $i = 1, 2, \dots, n$ ) is defined as :

$$E_i = \frac{\text{Var } T}{\text{Var } T_i}; i = 1, 2, \dots, n \quad \dots(15.12a)$$

Obviously  $E_i \leq 1$ ,  $i = 1, 2, \dots, n$ .

For example, in the normal samples, since sample mean  $\bar{x}$  is the most efficient estimator of  $\mu$  [c.f. Remark to Example 15-31], the efficiency  $E$  of  $Md$  for such samples, (for large  $n$ ), is :

$$E = \frac{V(\bar{x})}{V(Md)} = \frac{\sigma^2/n}{\pi\sigma^2/(2n)} = \frac{2}{\pi} = 0.637$$

**Example 15-7.** A random sample  $(X_1, X_2, X_3, X_4, X_5)$  of size 5 is drawn from a normal population with unknown mean  $\mu$ . Consider the following estimators to estimate  $\mu$  :

$$(i) \quad t_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

$$(ii) \quad t_2 = \frac{X_1 + X_2}{2} + X_3, \quad (iii) \quad t_3 = \frac{2X_1 + X_2 + \lambda X_3}{3}$$

where  $\lambda$  is such that  $t_3$  is an unbiased estimator of  $\mu$ .

Find  $\lambda$ . Are  $t_1$  and  $t_2$  unbiased? State giving reasons, the estimator which is best among  $t_1$ ,  $t_2$  and  $t_3$ .

**Solution.** We are given

$$E(X_i) = \mu, \text{Var}(X_i) = \sigma^2, \text{(say)}; \text{Cov}(X_i, X_j) = 0, (i \neq j = 1, 2, \dots, n)$$

...(\*)

$$(i) \quad E(t_1) = \frac{1}{5} \sum_{i=1}^5 E(X_i) = \frac{1}{5} \sum_{i=1}^5 \mu = \frac{1}{5} \cdot 5\mu = \mu$$

$\Rightarrow t_1$  is an unbiased estimator of  $\mu$ .

$$(ii) \quad E(t_2) = \frac{1}{2} E(X_1 + X_2) + E(X_3)$$

$$= \frac{1}{2} (\mu + \mu) + \mu \quad [\text{Using (*)}]$$

$$= 2\mu$$

$\Rightarrow t_2$  is not an unbiased estimator of  $\mu$ .

$$(iii) \quad E(t_3) = \mu$$

$$\Rightarrow \frac{1}{3} E(2X_1 + X_2 + \lambda X_3) = \mu$$

$$\Rightarrow 2E(X_1) + E(X_2) + \lambda E(X_3) = 3\mu$$

$$\Rightarrow 2\mu + \mu + \lambda\mu = 3\mu$$

$$\Rightarrow \lambda\mu = 0 \Rightarrow \lambda = 0$$

Using (\*), we get

$$V(t_1) = \frac{1}{25} [V(X_1) + V(X_2) + V(X_3) + V(X_4) + V(X_5)] = \frac{1}{5} \sigma^2$$

$$V(t_2) = \frac{1}{4} [V(X_1) + V(X_2)] + V(X_3) = \frac{1}{2} \sigma^2 + \sigma^2 = \frac{3}{2} \sigma^2$$

$$V(t_3) = \frac{1}{9} [4V(X_1) + V(X_2)] = \frac{1}{9} (4\sigma^2 + \sigma^2) = \frac{5}{9}\sigma^2 \quad (\because \lambda = 0)$$

Since  $V(t_1)$  is the least,  $t_1$  is the best estimator (in the sense of least variance) of  $\mu$ .

**Example 15.8.**  $X_1, X_2$ , and  $X_3$  is a random sample of size 3 from a population with mean value  $\mu$  and variance  $\sigma^2$ .  $T_1, T_2, T_3$  are the estimators used to estimate mean value  $\mu$ , where

$$T_1 = X_1 + X_2 - X_3, \quad T_2 = 2X_1 + 3X_3 - 4X_2, \quad \text{and} \quad T_3 = (\lambda X_1 + X_2 + X_3)/3$$

(i) Are  $T_1$  and  $T_2$  unbiased estimators?

(ii) Find the value of  $\lambda$  such that  $T_3$  is unbiased estimator for  $\mu$ .

(iii) With this value of  $\lambda$  is  $T_3$  a consistent estimator?

(iv) Which is the best estimator?

**Solution.** Since  $X_1, X_2, X_3$  is a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ ,

$$E(X_i) = \mu, \quad \text{Var}(X_i) = \sigma^2 \quad \text{and} \quad \text{Cov}(X_i, X_j) = 0, \quad (i \neq j = 1, 2, \dots, n) \quad \dots (*)$$

$$(i) \quad E(T_1) = E(X_1) + E(X_2) - E(X_3) = \mu + \mu - \mu = \mu$$

$\Rightarrow T_1$  is an unbiased estimator of  $\mu$

$$E(T_2) = 2E(X_1) + 3E(X_3) - 4E(X_2) = 2\mu + 3\mu - 4\mu = \mu$$

$\Rightarrow T_2$  is an unbiased estimator of  $\mu$ .

$$(ii) \quad \text{We are given :} \quad E(T_3) = \mu$$

$$\Rightarrow \frac{1}{3} [\lambda E(X_1) + E(X_2) + E(X_3)] = \mu$$

$$\Rightarrow \frac{1}{3} (\lambda\mu + \mu + \mu) = \mu \Rightarrow \lambda\mu + 2\mu = 3\mu \Rightarrow \lambda = 1.$$

$$(iii) \quad \text{With } \lambda = 1, \quad T_3 = \frac{1}{3}(X_1 + X_2 + X_3) = \bar{X}$$

Since sample mean is a consistent estimator of population mean  $\mu$ , by Weak Law of Large Numbers,  $T_3$  is a consistent estimator of  $\mu$ .

(iv) We have [on using (\*)] :

$$\text{Var}(T_1) = \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) = 3\sigma^2$$

$$\text{Var}(T_2) = 4 \text{Var}(X_1) + 9 \text{Var}(X_3) + 16 \text{Var}(X_2) = 29\sigma^2$$

$$\text{Var}(T_3) = \frac{1}{9} [\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)] = \frac{1}{3}\sigma^2 \quad (\because \lambda = 1)$$

Since  $\text{Var}(T_3)$  is minimum,  $T_3$  is the best estimator in the sense of minimum variance.

### 15.5.2. Minimum Variance Unbiased (M.V.U.) Estimators.

If a statistic  $T = T(x_1, x_2, \dots, x_n)$ , based on sample of size  $n$  is such that :

(i)  $T$  is unbiased for  $\gamma(\theta)$ , for all  $\theta \in \Theta$  and

(ii) It has the smallest variance among the class of all unbiased estimators of  $\gamma(\theta)$ .

then  $T$  is called the minimum variance unbiased estimator (MVUE) of  $\gamma(\theta)$ .

More precisely,  $T$  is MVUE of  $\gamma(\theta)$  if

$$E_\theta(T) = \gamma(\theta) \text{ for all } \theta \in \Theta \quad \dots(15.13)$$

$$\text{and} \quad \text{Var}_\theta(T) \leq \text{Var}_\theta(T') \text{ for all } \theta \in \Theta \quad \dots(15.14)$$

where  $T'$  is any other unbiased estimator of  $\gamma(\theta)$ .

We give below some important Theorems concerning MVU estimators.

**Theorem 15.3.** An M.V.U. is unique in the sense that if  $T_1$  and  $T_2$  are M.V.U. estimators for  $\gamma(\theta)$ , then  $T_1 = T_2$ , almost surely.

**Proof.** We are given that

$$\begin{aligned} E_\theta(T_1) &= E_\theta(T_2) = \gamma(\theta), \text{ for all } \theta \in \Theta \\ \text{and} \quad \text{Var}_\theta(T_1) &= \text{Var}_\theta(T_2) \text{ for all } \theta \in \Theta \end{aligned} \quad \left. \right\} \quad \dots(15.15)$$

Consider a new estimator

$$T = \frac{1}{2}(T_1 + T_2)$$

which is also unbiased since

$$E(T) = \frac{1}{2}[E(T_1) + E(T_2)] = \theta$$

$$\begin{aligned} \text{Var}(T) &= \text{Var}\left[\frac{1}{2}(T_1 + T_2)\right] = \frac{1}{4}\text{Var}(T_1 + T_2) [\because \text{Var}(C\bar{X}) = C^2 \text{Var}(X)] \\ &= \frac{1}{4}[\text{Var}(T_1) + \text{Var}(T_2) + 2\text{Cov}(T_1, T_2)] \\ &= \frac{1}{4}[\text{Var}(T_1) + \text{Var}(T_2) + 2\rho\sqrt{\text{Var}(T_1)\text{Var}(T_2)}] \\ &= \frac{1}{2}\text{Var}(T_1)(1 + \rho), \end{aligned} \quad \dots[(15.15)]$$

where  $\rho$  is Karl Pearson's co-efficient of correlation between  $T_1$  and  $T_2$ .

Since  $T_1$  is the MUV estimator,

$$\begin{aligned} \text{Var}(T) &\geq \text{Var}(T_1) \\ \Rightarrow \quad \frac{1}{2}\text{Var}(T_1)[1 + \rho] &\geq \text{Var}(T_1) \\ \Rightarrow \quad \frac{1}{2}(1 + \rho) &\geq 1, \text{ i.e., } \rho \geq 1 \end{aligned}$$

Since  $|\rho| \leq 1$ , we must have  $\rho = 1$ , i.e.,  $T_1$  and  $T_2$  must have a linear relation of the form :

$$T_1 = \alpha + \beta T_2, \quad \dots(15.16)$$

where  $\alpha$  and  $\beta$  are constants independent of  $x_1, x_2, \dots, x_n$  but may depend on  $\theta$ , i.e., we may have  $\alpha = \alpha(\theta)$  and  $\beta = \beta(\theta)$ .

Taking expectation of both sides in (15.16) and using (15.15), we get

$$\theta = \alpha + \beta\theta \quad \dots(15.17)$$

Also from (15.16), we get

$$\text{Var}(T_1) = \text{Var}(\alpha + \beta T_2) = \beta^2 \text{Var}(T_2)$$

$$\Rightarrow 1 = \beta^2 \Rightarrow \beta = \pm 1 \quad \dots[\text{From (15.15)}]$$

But since  $\rho(T_1, T_2) = +1$ , the coefficient of regression of  $T_1$  and  $T_2$  must be positive.

$$\therefore \beta = 1 \Rightarrow \alpha = 0 \quad [\text{From 15.17}]$$

Substituting in (15.16), we get  $T_1 = T_2$  as desired.

**Theorem 15.4.** Let  $T_1$  and  $T_2$  be unbiased estimators of  $\gamma(\theta)$  with efficiencies  $e_1$  and  $e_2$  respectively and  $\rho = \rho_\theta$  be the correlation coefficient between them. Then

$$\sqrt{e_1 e_2} - \sqrt{(1 - e_1)(1 - e_2)} \leq \rho \leq \sqrt{e_1 e_2} + \sqrt{(1 - e_1)(1 - e_2)}$$

**Proof.** Let  $T$  be the minimum variance unbiased estimator of  $\gamma(\theta)$ . Then we are given :

$$E_\theta(T_1) = \gamma(\theta) = E_\theta(T_2), \forall \theta \in \Theta \quad \dots(15.18)$$

$$\text{and} \quad e_1 = \frac{V_\theta(T)}{V_\theta(T_1)} = \frac{V}{V_1}, \text{ (say)} \Rightarrow V_1 = \frac{V}{e_1} \quad \dots(15.19)$$

$$e_2 = \frac{V_\theta(T)}{V_\theta(T_2)} = \frac{V}{V_2}, \text{ (say)} \Rightarrow V_2 = \frac{V}{e_2} \quad \dots(15.20)$$

Let us consider another estimator

$$T_3 = \lambda T_1 + \mu T_2 \quad \dots(15.21)$$

which is also unbiased estimator of  $\gamma(\theta)$ ,

$$\text{i.e.,} \quad E(T_3) = (\lambda + \mu) \gamma(\theta) = \gamma(\theta) \quad [\text{Using (15.18)}]$$

$$\Rightarrow \lambda + \mu = 1 \quad \dots(15.22)$$

$$\begin{aligned} V_\theta(T_3) &= V(\lambda T_1 + \mu T_2) \\ &= \lambda^2 V(T_1) + \mu^2 V(T_2) + 2\lambda\mu \text{Cov}(T_1, T_2) \\ &= V \left[ \frac{\lambda^2}{e_1} + \frac{\mu^2}{e_2} + 2 \cdot \frac{\lambda\mu\rho}{\sqrt{e_1 e_2}} \right] \quad [\text{Using (15.19) and (15.20)}] \end{aligned}$$

But  $V_\theta(T_3) \geq V$ , since  $V$  is the minimum variance.

$$\therefore \frac{\lambda^2}{e_1} + \frac{\mu^2}{e_2} + \frac{2\lambda\mu\rho}{\sqrt{e_1 e_2}} \geq 1 = (\lambda + \mu)^2 \quad [\text{Using (15.22)}]$$

$$\Rightarrow \left( \frac{1}{e_1} - 1 \right) \lambda^2 + \left( \frac{1}{e_2} - 1 \right) \mu^2 + 2\lambda\mu \left( \frac{\rho}{\sqrt{e_1 e_2}} - 1 \right) \geq 0$$

$$\Rightarrow \left( \frac{1}{e_1} - 1 \right) \left( \frac{\lambda}{\mu} \right)^2 + 2 \left( \frac{\rho}{\sqrt{e_1 e_2}} - 1 \right) \left( \frac{\lambda}{\mu} \right) + \left( \frac{1}{e_2} - 1 \right) \geq 0 \quad \dots(15.23)$$

which is quadratic expression in  $(\lambda/\mu)$ .

Note that :

$$e_i < 1 \Rightarrow \frac{1}{e_i} > 1 \Rightarrow \left( \frac{1}{e_i} - 1 \right) > 0; i = 1, 2$$

We know that

$$AX^2 + BX + C \geq 0 \quad \forall x, A > 0, C > 0;$$

if and only if

$$\text{Discriminant} = B^2 - 4AC \leq 0 \quad \dots(15.24)$$

Using (15.24), we get from (15.23) :

$$\begin{aligned} & \left( \frac{\rho}{\sqrt{e_1 e_2}} - 1 \right)^2 - \left( \frac{1}{e_1} - 1 \right) \left( \frac{1}{e_2} - 1 \right) \leq 0 \\ \Rightarrow & (\rho - \sqrt{e_1 e_2})^2 - (1 - e_1)(1 - e_2) \leq 0 \\ \Rightarrow & \rho^2 - 2\sqrt{e_1 e_2} \rho + (e_1 + e_2 - 1) \leq 0 \end{aligned}$$

This implies that  $\rho$  lies between the roots of the equation

$$\rho^2 - 2\sqrt{e_1 e_2} \rho + (e_1 + e_2 - 1) = 0$$

which are given by

$$\begin{aligned} & \frac{1}{2} [2\sqrt{e_1 e_2} \pm 2\sqrt{e_1 e_2 - (e_1 + e_2 - 1)}] \\ & = \sqrt{e_1 e_2} \pm \sqrt{(e_1 - 1)(e_2 - 1)} \end{aligned}$$

Hence we have :

$$\begin{aligned} & \sqrt{e_1 e_2} - \sqrt{(e_1 - 1)(e_2 - 1)} \leq \rho \leq \sqrt{e_1 e_2} + \sqrt{(e_1 - 1)(e_2 - 1)} \\ \Rightarrow & \sqrt{e_1 e_2} - \sqrt{(1 - e_1)(1 - e_2)} \leq \rho \leq \sqrt{e_1 e_2} + \sqrt{(1 - e_1)(1 - e_2)} \end{aligned} \quad \dots(15.25)$$

**Corollary.** If we take  $e_1 = 1$  and  $e_2 = e$  in (15.25), we get

$$\sqrt{e} \leq \rho \leq \sqrt{e} \Rightarrow \rho = \sqrt{e}$$

This leads to the following important result, which we state in the form of a theorem.

**Theorem 15.5.** If  $T_1$  is an MVU estimator of  $\gamma(\theta)$ ,  $\theta \in \Theta$  and  $T_2$  is any other unbiased estimator of  $\gamma(\theta)$  with efficiency  $e = e_\theta$ , then the correlation coefficient between  $T_1$  and  $T_2$  is given by

$$\rho = \sqrt{e} \quad \text{i.e., } \rho_\theta = \sqrt{e_\theta}, \forall \theta \in \Theta.$$

For an alternate proof, see Examples 15.9 and 15.10.

**Theorem 15.6.** If  $T_1$  is an MVUE of  $\gamma(\theta)$  and  $T_2$  is any other unbiased estimator of  $\gamma(\theta)$  with efficiency  $e < 1$ , then no unbiased linear combination of  $T_1$  and  $T_2$  can be an MVUE of  $\gamma(\theta)$ .

**Proof.** A linear combination

$$T = l_1 T_1 + l_2 T_2 \quad \dots(15.27)$$

will be unbiased estimator of  $\gamma(\theta)$  if

$$E(T) = l_1 E(T_1) + l_2 E(T_2) = \gamma(\theta), \text{ for all } \theta \in \Theta$$

$$\Rightarrow l_1 + l_2 = 1 \quad \dots(15.27a)$$

since we are given  $E(T_1) = E(T_2) = \gamma(\theta)$ .

We have

$$e = \frac{\text{Var}(T_1)}{\text{Var}(T_2)} \Rightarrow \text{Var} T_2 = \frac{\text{Var} T_1}{e} \quad \dots(15.28)$$

and

$$\rho = \rho(T_1, T_2) = \sqrt{e} \quad [\text{c.f. Theorem 15.5}]$$

From (15.27), on using (15.28), we get

$$\begin{aligned} \text{Var } T &= l_1^2 \text{Var}(T_1) + l_2^2 \text{Var}(T_2) + 2l_1 l_2 \text{Cov}(T_1, T_2) \\ &= l_1^2 \text{Var}(T_1) + l_2^2 \text{Var}(T_2) + 2l_1 l_2 \rho \sqrt{\text{Var}(T_1) \text{Var}(T_2)} \\ &= \text{Var}(T_1) \left[ l_1^2 + \frac{l_2^2}{e} + 2l_1 l_2 \frac{\rho}{\sqrt{e}} \right] \\ &= \text{Var}(T_1) \left[ l_1^2 + 2l_1 l_2 + \frac{l_2^2}{e} \right] \quad (\because \rho \sqrt{e}) \\ &\geq \text{Var} T_1 [l_1^2 + 2l_1 l_2 + l_2^2] \quad \left( \because 0 < e \leq 1 \Rightarrow \frac{1}{e} \geq 1 \right) \\ &= \text{Var} T_1 (l_1 + l_2)^2 \\ &= \text{Var}(T_1) \quad [\text{From (15.27a)}] \end{aligned}$$

$\Rightarrow T$  cannot be an MVU estimator.

**Example 15.9.** If  $T_1$  and  $T_2$  be two unbiased estimators of  $\gamma(\theta)$  with variances  $\sigma_1^2, \sigma_2^2$  and correlation  $\rho$ , what is the best unbiased linear combination of  $T_1$  and  $T_2$  and what is the variance of such a combination?

[Delhi Univ.B.Sc. (Stat. Hons.), 1990]

**Solution.** Let  $T_1$  and  $T_2$  be two unbiased estimators of  $\gamma(\theta)$ .

$$\therefore E(T_1) = E(T_2) = \gamma(\theta) \quad \dots(1)$$

Let  $T$  be a linear combination of  $T_1$  and  $T_2$  given by

$$T = l_1 T_1 + l_2 T_2 \quad (*)$$

where  $l_1, l_2$  are arbitrary constants.

$$E(T) = l_1 E(T_1) + l_2 E(T_2) = (l_1 + l_2) \gamma(\theta) \quad [\text{From (1)}]$$

$\therefore T$  is also an unbiased estimator of  $\gamma(\theta)$  if and only if

$$l_1 + l_2 = 1 \quad \dots(2)$$

$$\begin{aligned} \text{Now } V(T) &= V(l_1 T_1 + l_2 T_2) \\ &= l_1^2 V(T_1) + l_2^2 V(T_2) + 2l_1 l_2 \text{Cov}(T_1, T_2) \\ &= l_1^2 \sigma_1^2 + l_2^2 \sigma_2^2 + 2l_1 l_2 \rho \sigma_1 \sigma_2 \quad \dots(3) \end{aligned}$$

We want the minimum value of (3) for variations in  $l_1$  and  $l_2$ , subject to the condition (2).

$$\therefore \frac{\partial}{\partial l_1} V(T) = 0 = l_1 \sigma_1^2 + l_2 \rho \sigma_1 \sigma_2$$

$$\frac{\partial}{\partial l_2} V(T) = 0 = l_2 \sigma_2^2 + l_1 \rho \sigma_1 \sigma_2$$

Substracting, we get

$$\begin{aligned} l_1(\sigma_1^2 - \rho \sigma_1 \sigma_2) &= l_2(\sigma_2^2 - \rho \sigma_1 \sigma_2) \\ \Rightarrow \frac{l_1}{\sigma_2^2 - \rho \sigma_1 \sigma_2} &= \frac{l_2}{\sigma_1^2 - \rho \sigma_1 \sigma_2} = \frac{l_1 + l_2}{\sigma_1^2 + \sigma_2^2 - 2\rho \sigma_1 \sigma_2} \\ &= \frac{1}{\sigma_1^2 + \sigma_2^2 - 2\rho \sigma_1 \sigma_2} \quad [\text{From (2)}] \end{aligned}$$

$$\therefore l_1 = \frac{\sigma_2^2 - \rho \sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho \sigma_1 \sigma_2} \text{ and } l_2 = \frac{\sigma_1^2 - \rho \sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho \sigma_1 \sigma_2} \quad \dots(4)$$

With these values of  $l_1$  and  $l_2$ ,  $T$  given by (\*) is the best unbiased linear combination of  $T_1$  and  $T_2$  and its variance is given by (3).

**Example 15.10.** Suppose  $T_1$  in the above example is an unbiased minimum variance estimate and  $T_2$  is any other unbiased estimate with variance  $\sigma^2/e$ . Then prove that the correlation between  $T_1$  and  $T_2$  is  $\sqrt{e}$ .

**Solution.** The coefficients of the best linear unbiased combination of  $T_1$  and  $T_2$ , given by (\*) in Example 15.9 are given by (4).

We are given that  $\sigma_1^2 = V(T_1) = \sigma^2$

$$\text{and } e = \frac{V(T_1)}{V(T_2)} = \frac{\sigma^2}{V(T_2)} \Rightarrow V(T_2) = \sigma_2^2 = \sigma^2/e$$

Substituting in (4) of Example 15.9, we get

$$\left. \begin{array}{l} l_1 = \frac{1 - \rho \sqrt{e}}{D} \\ l_2 = \frac{e - \rho \sqrt{e}}{D} \end{array} \right\}, \text{ where } D = 1 + e - 2\rho \sqrt{e} \quad \dots(5)$$

Hence from (\*), the unbiased statistic is

$$T = \frac{[(1 - \rho \sqrt{e}) T_1 + (e - \rho \sqrt{e}) T_2]}{D}$$

and from (3) the minimum variance is :

$$\begin{aligned} V(T) &= \frac{1}{D^2} \left[ (1 - \rho \sqrt{e})^2 \sigma^2 + (e - \rho \sqrt{e})^2 \frac{\sigma^2}{e} + 2(1 - \rho \sqrt{e})(e - \rho \sqrt{e}) \cdot \rho \cdot \sigma \cdot \sigma / \sqrt{e} \right] \\ &= \frac{\sigma^2}{D^2} \left[ (1 + \rho^2 e - 2\rho \sqrt{e}) + \frac{1}{e} (e^2 + \rho^2 e - 2\rho e \sqrt{e}) + 2(1 - \rho \sqrt{e})(\sqrt{e} - \rho) \rho \right] \\ &= \frac{\sigma^2}{D^2} \left[ 1 + \rho^2 e - 2\rho \sqrt{e} + e + \rho^2 - 2\rho \sqrt{e} + 2(\rho \sqrt{e} - \rho^2 e - \rho^2 + \rho^3 \sqrt{e}) \right] \end{aligned}$$

$$\begin{aligned}
 &= \frac{\sigma^2}{D^2} \left[ 1 - \rho^2 e + e - \rho^2 - 2\rho \sqrt{e} + 2\rho^3 \sqrt{e} \right] \\
 &= \frac{\sigma^2}{D^2} \left[ (1 + e - 2\rho \sqrt{e}) - \rho^2(e + 1 - 2\rho \sqrt{e}) \right] \\
 &= \frac{\sigma^2(1 - \rho^2)(1 + e - 2\rho \sqrt{e})}{(1 + e - 2\rho \sqrt{e})^2} = \frac{\sigma^2(1 - \rho^2)}{1 + e - 2\rho \sqrt{e}} \\
 &= \frac{\sigma^2(1 - \rho^2)}{(1 - \rho^2) + (\sqrt{e} - \rho)^2} \\
 \therefore \quad \frac{V(T)}{\sigma^2} &= \frac{1 - \rho^2}{(1 - \rho^2) + (\sqrt{e} - \rho)^2} \leq 1 \quad \dots(6)
 \end{aligned}$$

Since  $T_1$  is the most efficient estimator,

$$V(T) \leq \sigma^2 \Rightarrow \frac{V(T)}{\sigma^2} \geq 1 \quad \dots(7)$$

From (6) and (7), we get

$$\begin{aligned}
 \frac{V(T)}{\sigma^2} &= 1, \text{ i.e., } \frac{1 - \rho^2}{(1 - \rho^2) + (\sqrt{e} - \rho)^2} = 1 \\
 \Rightarrow (\sqrt{e} - \rho)^2 &= 0 \Rightarrow \rho = \sqrt{e}
 \end{aligned}$$

**Aliter.** From (5) onwards. Since  $T_1$  is given to be the most efficient estimator, it cannot be improved upon (c.f. Theorem 15-6). Hence, in order that  $T$  defined (\*) is minimum variance unbiased estimator we must have

$$\left. \begin{array}{l} l_1 = 1 \\ l_2 = 0 \end{array} \right\} \Rightarrow \rho = \sqrt{e} \quad \dots[\text{From (5)}]$$

**Remark.** This problem leads to the following very important result :

"The correlation coefficient between a most efficient estimator and any other estimator with efficiency  $e$  is  $\sqrt{e}$ ."

**Example 15-11.** (a) Show that if a most efficient estimator  $A$  and a less efficient estimator  $B$  with efficiency  $e$  tend to joint normality for large samples,  $B - A$  tends to zero correlation with  $A$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

(b) Show that the error in  $B$  may be regarded as composed (for large samples) of two parts which are independent, the error in  $A$  and the error in  $(B - A)$ .

(c) Show further that

$$V(B - A) = \left( \frac{1}{e} - 1 \right) V(A)$$

**Solution.** (a) We have to prove that

$$r[A, (B - A)] = 0 \Rightarrow \text{Cov}(A, B - A) = 0$$

$$\text{Cov}[A, (B - A)] = \text{Cov}(A, B) - V(A) = \rho\sigma_A\sigma_B - \sigma_A^2,$$

where  $\rho$  is the correlation coefficient between  $A$  and  $B$ .

$$\text{If we take } \sigma_A = \sigma, \text{ then } \sigma_B = \frac{\sigma}{\sqrt{e}} \text{ and } \rho = \sqrt{e} \quad (\text{c.f. Theorem 15-5})$$

$$\therefore \text{Cov}(A, B - A) = \sqrt{e} \cdot \sigma \cdot \frac{\sigma}{\sqrt{e}} - \sigma^2 = 0$$

Hence  $(B - A)$  has zero correlation with  $A$ .

(b) We have  $B = A + (B - A)$

$$\begin{aligned} \therefore V(B) &= V[A + (B - A)] = V(A) + V(B - A) + 2 \text{Cov}(A, B - A) \\ &= V(A) + V(B - A) \quad [\text{Using part (a)}] \end{aligned}$$

$$\Rightarrow \text{Error in } B = \text{Error in } A + \text{Error in } (B - A)$$

and since  $A$  and  $(B - A)$  are independent, [c.f. part (a) viz.,  $r(A, B - A) = 0$  and  $A$  and  $B$  tend to joint normality], the result follows.

$$\begin{aligned} (c) \quad V(A - B) &= V(A) + V(B) - 2 \text{Cov}(A, B) \\ &= \sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B \\ &= \sigma^2 + \frac{\sigma^2}{e} - 2\sqrt{e}\cdot\sigma\cdot\frac{\sigma}{\sqrt{e}} \\ &= \frac{\sigma^2}{e} - \sigma^2 = \left(\frac{1}{e} - 1\right)\sigma^2 \end{aligned}$$

**Example 15-12.** If  $T_1$  and  $T_2$  are two unbiased estimators of  $\gamma(\theta)$ , having the same variance and  $\rho$  is the correlation between them, then show that  $\rho \geq 2e - 1$ , where  $e$  is the efficiency of each estimator.

**Solution.** Let  $T$  be MVUE of  $\gamma(\theta)$ . Then, since  $V(T_1) = V(T_2)$ , the efficiency  $e$  of each estimator is given by :

$$e = \frac{V(T)}{V(T_1)} = \frac{V(T)}{V(T_2)} \quad \dots(*)$$

Consider another unbiased estimator of  $\gamma(\theta)$  viz.,

$$\begin{aligned} T_3 &= \frac{1}{2}(T_1 + T_2) \\ \Rightarrow V(T_3) &= \frac{1}{4}[V(T_1) + V(T_2) + 2 \text{Cov}(T_1, T_2)] \\ &= \frac{1}{4} \left[ \frac{V(T)}{e} + \frac{V(T)}{e} + 2\rho \sqrt{\frac{V(T)}{e} \cdot \frac{V(T)}{e}} \right] \\ &= \frac{V(T)}{4e} [1 + 1 + 2\rho] = \frac{(1 + \rho)V(T)}{2e} \end{aligned}$$

Since  $V(T)$  is the minimum variance,

$$V(T_3) = \frac{(1 + \rho) \cdot V(T)}{2e} \geq V(T)$$

$$\Rightarrow 1 + p \geq 2e \Rightarrow p \geq (2e - 1).$$

**Aliter.** Deduction From (15-25). If  $T_1$  and  $T_2$  have same variances/efficiencies i.e.,  $e_1 = e_2 = e$ , (say) then (15-25) gives

$$e - (1 - e) \leq p \leq e + (1 - e) \Rightarrow p \geq 2e - 1.$$

**15-6. Sufficiency.** An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter. More precisely, if  $T = t(x_1, x_2, \dots, x_n)$  is an estimator of a parameter  $\theta$ , based on a sample  $x_1, x_2, \dots, x_n$  of size  $n$  from the population with density  $f(x, \theta)$  such that the conditional distribution of  $x_1, x_2, \dots, x_n$  given  $T$ , is independent of  $\theta$ , then  $T$  is sufficient estimator for  $\theta$ .

**Illustration.** Let  $x_1, x_2, \dots, x_n$  be a random sample from a Bernoulli population with parameter ' $p$ ',  $0 < p < 1$ , i.e.,

$$x_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } q = (1 - p) \end{cases}$$

$$\text{Then } T = t(x_1, x_2, \dots, x_n) = x_1 + x_2 + \dots + x_n \sim B(n, p)$$

$$\therefore P(T = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The conditional distribution of  $(x_1, x_2, \dots, x_n)$  given  $T$  is

$$P[x_1 \cap x_2 \cap \dots \cap x_n | T = k] = \frac{P[x_1 \cap x_2 \cap \dots \cap x_n \cap T = k]}{P(T = k)}$$

$$= \begin{cases} \frac{p^k (1-p)^{n-k}}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{1}{\binom{n}{k}} \\ 0, \text{ if } \sum_{i=1}^n x_i \neq k \end{cases}$$

Since this does not depend on ' $p$ ',  $T = \sum_{i=1}^n x_i$ , is sufficient for ' $p$ '.

**Theorem 15-7. Factorization Theorem (Neyman).** The necessary and sufficient condition for a distribution to admit sufficient statistic is provided by the 'factorization theorem' due to Neyman.

**Statement**  $T = t(x)$  is sufficient for  $\theta$  if and only if the joint density function  $L$  (say), of the sample values can be expressed in the form

$$L = g_\theta[t(x)].h(x) \quad \dots(15-29)$$

where (as indicated)  $g_\theta[t(x)]$  depends on  $\theta$  and  $x$  only through the value of  $t(x)$  and  $h(x)$  is independent of  $\theta$ .

**Remarks 1.** It should be clearly understood that by 'a function independent of  $\theta$ ' we not only mean that it does not involve  $\theta$  but also that its domain does not contain  $\theta$ . For example, the function

$$f(x) = \frac{1}{2a}, a - \theta < x < a + \theta ; -\infty < \theta < \infty$$

depends on  $\theta$ .

2. It should be noted that the original sample  $X = (X_1, X_2, \dots, X_n)$ , is always a sufficient statistic.

3. The most general form of the distributions admitting sufficient statistic is *Koopman's form* and is given by

$$L = L(x, \theta) = g(x).h(\theta). \exp \{a(\theta)\psi(x)\} \quad \dots(15.30)$$

where  $h(\theta)$  and  $a(\theta)$  are functions of the parameter  $\theta$  only and  $g(x)$  and  $\psi(x)$  are the functions of the sample observations only.

Equation (15.30) represents the famous *exponential family of distributions*, of which most of the common distributions like the binomial, the Poisson and the normal with unknown mean and variance, are the members.

#### 4. Invariance Property of Sufficient Estimator.

If  $T$  is a sufficient estimator for the parameter  $\theta$  and if  $\psi(T)$  is a one to one function of  $T$ , then  $\psi(T)$  is sufficient for  $\psi(\theta)$ .

5. Fisher-Neyman Criterion. A statistic  $t_1 = t_1(x_1, x_2, \dots, x_n)$  is sufficient estimator of parameter  $\theta$  if and only if the likelihood function (joint p.d.f. of the sample) can be expressed as :

$$\begin{aligned} L &= \prod_{i=1}^n f(x_i, \theta) \\ &= g_1(t_1, \theta). k(x_1, x_2, \dots, x_n) \end{aligned} \quad \dots(15.31)$$

where  $g_1(t_1, \theta)$  is the p.d.f. of statistic  $t_1$  and  $k(x_1, x_2, \dots, x_n)$  is a function of sample observations only independent of  $\theta$ .

Note that this method requires the working out of the p.d.f. (p.m.f.) of the statistic  $t_1 = t(x_1, x_2, \dots, x_n)$ , which is not always easy.

**Example 15.13.** Let  $x_1, x_2, \dots, x_n$  be a random sample from a uniform population on  $[0, \theta]$ . Find a sufficient estimator for  $\theta$ .

[Madras Univ. B.Sc., Oct. 1992]

**Solution.** We are given

$$f_\theta(x_i) = \begin{cases} \frac{1}{\theta}, & 0 \leq x_i \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

Let  $k(a, b) = \begin{cases} 1, & \text{if } a \leq b \\ 0, & \text{if } a > b \end{cases}$

Then  $f_\theta(x_i) = \frac{k(0, x_i) k(x_i, \theta)}{\theta}$ ,

$$L = \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \left[ \frac{k(0, x_i) k(x_i, \theta)}{\theta} \right]$$

$$\begin{aligned} &\frac{k(0, \min_{1 \leq i \leq n} x_i) \cdot k(\max_{1 \leq i \leq n} x_i, \theta)}{\theta^n} = g_\theta[t(x)] h(x) \end{aligned}$$

where  $g_{\theta}[t(x)] = \frac{k(t(x), \theta)}{\theta^n}$ ,  $t(x) = \max_{1 \leq i \leq n} x_i$  and  $h(x) = k(0, \min_{1 \leq i \leq n} x_i)$

Hence by Factorization Theorem,  $T = \max_{1 \leq i \leq n} x_i$ , is sufficient statistic for  $\theta$ .

**Aliter.** We have  $L = \prod_{i=1}^n f(x_i, \theta) = \frac{1}{\theta^n} ; 0 < x_i < \theta$  ... (i)

If  $t = \max(x_1, x_2, \dots, x_n) = x_{(n)}$ , then p.d.f. of  $T$  is given by :  
 $g(t, \theta) = n [F(x_{(n)})]^{n-1} \cdot f(x_{(n)})$  ... (ii)

We have  $F(x) = P(X \leq x) = \int_0^x f(x, \theta) dx = \int_0^x \frac{1}{\theta} \cdot dx = \frac{x}{\theta}$

$$\therefore g(t, \theta) = n \left[ \frac{x_{(n)}}{\theta} \right]^{n-1} \left( \frac{1}{\theta} \right) \quad [\text{From (ii)}]$$

$$= \frac{n}{\theta^n} [x_{(n)}]^{n-1}$$

Rewriting (i), we get

$$L = \frac{n [x_{(n)}]^{n-1}}{\theta^n} \cdot \frac{1}{n [x_{(n)}]^{n-1}}$$

$$= g(t, \theta) \cdot h(x_1, x_2, \dots, x_n)$$

Hence by Fisher-Neyman criterion, the statistic  $t = x_{(n)}$ , is sufficient estimator for  $\theta$ .

**Example 15-14.** Let  $x_1, x_2, \dots, x_n$  be a random sample from  $N(\mu, \sigma^2)$  population. Find sufficient estimators for  $\mu$  and  $\sigma^2$ .

**Solution.** Let us write

$$\theta = (\mu, \sigma^2); -\infty < \mu < \infty, 0 < \sigma^2 < \infty$$

Then

$$L = \prod_{i=1}^n f_{\theta}(x_i) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right]$$

$$= \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^n x_i^2 - 2\mu \sum x_i + n\mu^2 \right) \right]$$

$$= g_{\theta}[t(x)]. h(x)$$

where

$$g_{\theta}[t(x)] = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \{ t_2(x) - 2\mu t_1(x) + n\mu^2 \} \right],$$

$$t(x) = (t_1(x), t_2(x)) = (\sum x_i, \sum x_i^2) \text{ and } h(x) = 1$$

Thus  $t(x) = \sum x_i$  is sufficient for  $\mu$  and  $t_2(x) = \sum x_i^2$ , is sufficient for  $\sigma^2$ .

**Example 15.15.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with p.d.f.

$$f(x, \theta) = e^{-(x-\theta)}, \theta < x < \infty; -\infty < \theta < \infty$$

Obtain sufficient statistic for  $\theta$ .

**Solution.** Here

$$\begin{aligned} L &= \prod_{i=1}^n f(x_i, \theta) = \prod_{i=1}^n [e^{-(x_i-\theta)}] \\ &= \exp \left[ - \sum_{i=1}^n x_i \right] \times \exp(n\theta) \end{aligned} \quad \dots (*)$$

Let  $Y_1, Y_2, \dots, Y_n$  denote the order statistics of the random sample such that  $Y_1 < Y_2 < \dots < Y_n$ . The p.d.f. of the smallest observation  $Y_1$  is given by

$$g_1(y_1, \theta) = n[1 - F(y_1)]^{n-1} f(y_1, \theta)$$

where  $F(\cdot)$  is the distribution function corresponding to p.d.f.  $f(\cdot)$ .

$$\text{Now } F(x) = \int_{-\theta}^x e^{-(x-\theta)} dx = \left| \frac{e^{-(x-\theta)}}{-1} \right|_{-\theta}^x = 1 - e^{-(x-\theta)}$$

$$\begin{aligned} \therefore g_1(y_1, \theta) &= n [e^{-(y_1-\theta)}]^{n-1} \cdot e^{-(y_1-\theta)} \\ &= n e^{-n(y_1-\theta)}, \theta < y_1 < \infty \\ &= 0, \text{ otherwise} \end{aligned}$$

Thus the likelihood function of  $X_1, X_2, \dots, X_n$  may be expressed as

$$\begin{aligned} L &= e^{n\theta} \exp \left( - \sum_{i=1}^n x_i \right) \\ &= n \exp \{ -n(y_1 - \theta) \} \left[ \frac{\exp \left( - \sum_{i=1}^n x_i \right)}{n \exp (-ny_1)} \right] \\ &= g_1(\min x_i, \theta) \left[ \frac{\exp \left( - \sum_{i=1}^n x_i \right)}{n \exp (-n \min x_i)} \right] \end{aligned}$$

Hence by Fisher-Neyman criterion, the first order statistic

$Y_1 = \min(X_1, X_2, \dots, X_n)$  is a sufficient statistic for  $\theta$ .

**Example 15.16.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with p.d.f.

$$f(x, \theta) = \theta x^{\theta-1}; 0 < x < 1, \theta > 0.$$

Show that  $t_1 = \prod_{i=1}^n X_i$ , is sufficient for  $\theta$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1988; Agra Univ. B.Sc., 1992]

$$\begin{aligned}
 \text{Solution. } L(x, \theta) &= \prod_{i=1}^n f(x_i, \theta) = \theta^n \prod_{i=1}^n (x_i^{\theta-1}) \\
 &= \theta^n \left( \prod_{i=1}^n x_i \right)^{\theta} \cdot \frac{1}{\left( \prod_{i=1}^n x_i \right)} \\
 &= g(t_1, \theta) \cdot h(x_1, x_2, \dots, x_n), \text{ (say).}
 \end{aligned}$$

Hence by Factorisation Theorem,

$$t_1 = \prod_{i=1}^n X_i \text{ is sufficient estimator for } \theta.$$

**Example 15.17.** Let  $X_1, X_2, \dots, X_n$  be a random sample from Cauchy population :

$$f(x, \theta) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}; -\infty < x < \infty, -\infty < \theta < \infty.$$

Examine if there exists a sufficient statistic for  $\theta$ .

$$\begin{aligned}
 \text{Solution. } L(x, \theta) &= \prod_{i=1}^n f(x_i, \theta) = \frac{1}{\pi^n} \cdot \prod_{i=1}^n \left[ \frac{1}{1 + (x_i - \theta)^2} \right] \\
 &\neq g(t_1, \theta) \cdot h(x_1, x_2, \dots, x_n).
 \end{aligned}$$

Hence by Factorisation Theorem, there is no single statistic, which alone, is sufficient estimator of  $\theta$ .

However,

$$L(x, \theta) = k_1(X_1, X_2, \dots, X_n, \theta) \cdot k_2(X_1, X_2, \dots, X_n)$$

$\Rightarrow$  The whole set  $(X_1, X_2, \dots, X_n)$  is jointly sufficient for  $\theta$ .

### 15.7. Cramer-Rao Inequality

**Theorem 15.8.** If  $t$  is an unbiased estimator for  $\gamma(\theta)$ , a function of parameter  $\theta$ , then

$$Var(t) \geq \frac{\left[ \frac{d}{d\theta} \cdot \gamma(\theta) \right]^2}{E \left[ \frac{\partial}{\partial \theta} \log L \right]^2} = \frac{[\gamma'(\theta)]^2}{I(\theta)} \quad \dots(15.32)$$

where  $I(\theta)$  is the information on  $\theta$ , supplied by the sample.

In other words, Cramer-Rao inequality provides a lower bound  $[\gamma'(\theta)]^2/I(\theta)$ , to the variance of an unbiased estimator of  $\gamma(\theta)$ .

**Proof.** In proving this result, we assume that there is only a single parameter  $\theta$  which is unknown. We also take the case of continuous r.v. The case of discrete random variables can be dealt with similarly on replacing the multiple integrals by appropriate multiple sums.

We further make the following assumptions, which are known as the *Regularity conditions for Cramer-Rao Inequality*.

(1) The parameter space  $\Theta$  is a non-degenerate open interval on the real line  $R^1 (-\infty, \infty)$ .

(2) For almost all  $x = (x_1, x_2, \dots, x_n)$ , and for all  $\theta \in \Theta$ ,  $\frac{\partial}{\partial \theta} L(x, \theta)$  exists, the exceptional set, if any, is independent of  $\theta$ .

(3) The range of integration is independent of the parameter  $\theta$ , so that  $f(x, \theta)$  is differentiable under integral sign.

If range is not independent of  $\theta$  and  $f$  is zero at the extremes of the range, i.e.,  $f(a, \theta) = 0 = f(b, \theta)$ , then

$$\begin{aligned}\frac{\partial}{\partial \theta} \int_a^b f dx &= \int_a^b \frac{\partial f}{\partial \theta} dx - f(a, \theta) \frac{\partial a}{\partial \theta} + f(b, \theta) \frac{\partial b}{\partial \theta} \\ \Rightarrow \quad \frac{\partial}{\partial \theta} \int_a^b f dx &= \int_a^b \frac{\partial f}{\partial \theta} dx, \text{ since } f(a, \theta) = 0 = f(b, \theta)\end{aligned}$$

(4) The conditions of uniform convergence of integrals are satisfied so that differentiation under the integral sign is valid.

$$(5) I(\theta) = E \left[ \left\{ \frac{\partial}{\partial \theta} \log L(x, \theta) \right\}^2 \right], \text{ exists and is positive for all } \theta \in \Theta.$$

Let  $X$  be a r.v. following the p.d.f.  $f(x, \theta)$  and let  $L$  be the likelihood function of the random sample  $(x_1, x_2, \dots, x_n)$  from this population. Then

$$L = L(x, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

Since  $L$  is the joint p.d.f. of  $(x_1, x_2, \dots, x_n)$ ,

$$\int L(x, \theta) dx = 1,$$

$$\text{where } \int dx = \int \int \dots \int dx_1 dx_2 \dots dx_n.$$

Differentiating w.r. to  $\theta$  and using regularity conditions given above, we get :

$$\begin{aligned}\int \frac{\partial}{\partial \theta} L dx = 0 &\Rightarrow \int \left( \frac{\partial}{\partial \theta} \log L \right) L dx = 0 \\ \Rightarrow \quad E \left( \frac{\partial}{\partial \theta} \log L \right) &= 0 \quad \dots(15-33)\end{aligned}$$

Let  $t = t(x_1, x_2, \dots, x_n)$  be an unbiased estimator of  $\gamma(\theta)$  such that

$$E(t) = \gamma(\theta) \Rightarrow \int t \cdot L dx = \gamma(\theta) \quad \dots(15-34)$$

Differentiating w.r. to  $\theta$ , we get

$$\int t \cdot \frac{\partial L}{\partial \theta} dx = \gamma'(\theta) \Rightarrow \int t \left( \frac{\partial}{\partial \theta} \log L \right) L dx = \gamma'(\theta)$$

$$\Rightarrow E\left(t \cdot \frac{\partial}{\partial \theta} \log L\right) = \gamma'(\theta) \quad \dots(15.35)$$

$$\text{Cov}\left(t, \frac{\partial}{\partial \theta} \log L\right) = E\left[t \cdot \frac{\partial}{\partial \theta} \log L\right] - E(t) \cdot E\left(\frac{\partial}{\partial \theta} \log L\right) \\ = \gamma'(\theta) \quad [\text{From (15.33) and (15.35)}]$$

We have :

$$\begin{aligned} [r(X, Y)]^2 &\leq 1 \Rightarrow [\text{Cov}(X, Y)]^2 \leq \text{Var}(X) \cdot \text{Var}(Y) \\ \therefore \left[\text{Cov}\left(t, \frac{\partial}{\partial \theta} \log L\right)\right]^2 &\leq \text{Var}(t) \cdot \text{Var}\left(\frac{\partial}{\partial \theta} \log L\right) \\ \Rightarrow [\gamma'(\theta)]^2 &\leq \text{Var}(t) \left[ E\left(\frac{\partial}{\partial \theta} \log L\right)^2 - \left\{ E\left(\frac{\partial}{\partial \theta} \log L\right) \right\}^2 \right] \\ \Rightarrow [\gamma'(\theta)]^2 &\leq \text{Var}(t) \cdot E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right] \quad [\text{Using (15.33)}] \dots(15.36) \\ \Rightarrow \text{Var}(t) &\geq \frac{[\gamma'(\theta)]^2}{E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right]} \quad \dots(15.36a) \end{aligned}$$

which is Cramer-Rao Inequality.

**Corollary.** If  $t$  is an unbiased estimator of parameter  $\theta$  i.e.,

$$E(t) = \theta \Rightarrow \gamma(\theta) = \theta \Rightarrow \gamma'(\theta) = 1,$$

then from (15.36a), we get

$$\text{Var}(t) \geq \frac{1}{E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right]} = \frac{1}{I(\theta)} \quad \dots(15.37)$$

where

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right] \quad \dots(15.37a)$$

is called by R.A. Fisher as the *amount of information* on  $\theta$  supplied by the sample.  $(x_1, x_2, \dots, x_n)$  and its reciprocal  $1/I(\theta)$ , as the *information limit* to the variance of estimator  $t = t(x_1, x_2, \dots, x_n)$ .

**Remarks.** 1. An unbiased estimator  $t$  of  $\gamma(\theta)$  for which Cramér-Rao lower bound in (15.32) is attained is called a *minimum variance bound (MVB) estimator*.

2. We have :

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right] = -E\left[\frac{\partial^2}{\partial \theta^2} \log L\right] \quad \dots(15.38)$$

$$\text{and} \quad I(\theta) = n \left[ \frac{\partial}{\partial \theta} \log f(x, \theta) \right]^2 = -n \left[ \frac{\partial^2}{\partial \theta^2} \log f \right] \quad \dots(15.38a)$$

**Proof.** We have proved in (15.33),

$$E\left(\frac{\partial}{\partial \theta} \log L\right) = 0 \quad \dots(*)$$

Also

$$\begin{aligned} \left(\frac{\partial^2}{\partial \theta^2} \log L\right)_L &= \frac{\partial}{\partial \theta} \left[ \left(\frac{\partial}{\partial \theta} \log L\right) \cdot L \right] - \left(\frac{\partial}{\partial \theta} \log L\right) \cdot \frac{\partial L}{\partial \theta} \\ &= \frac{\partial}{\partial \theta} \left[ \left(\frac{\partial}{\partial \theta} \log L\right) \cdot L \right] - \left(\frac{\partial}{\partial \theta} \log L\right)^2 \cdot L \end{aligned}$$

Integrating both sides w.r. to  $x = (x_1, x_2, \dots, x_n)$ , we get

$$\begin{aligned} E\left(\frac{\partial^2}{\partial \theta^2} \log L\right) &= \frac{\partial}{\partial \theta} \cdot E\left(\frac{\partial}{\partial \theta} \log L\right) - E\left(\frac{\partial}{\partial \theta} \log L\right)^2 \\ &= -E\left(\frac{\partial}{\partial \theta} \log L\right)^2 \quad [\text{Using } (*)] \\ \Rightarrow I(\theta) &= E\left(\frac{\partial}{\partial \theta} \log L\right)^2 = -E\left(\frac{\partial^2}{\partial \theta^2} \log L\right), \end{aligned}$$

a form which is more convenient to use in practice.

$$\begin{aligned} \text{Also } I(\theta) &= E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right] = E\left[\sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(x_i, \theta)\right]^2 \\ &= E\left[\sum_{i=1}^n \left\{\frac{\partial}{\partial \theta} \log f(x_i, \theta)\right\}^2\right] \\ &\quad + \sum_{i \neq j=1}^n \left\{ \left(\frac{\partial}{\partial \theta} \log f(x_i, \theta)\right) \cdot \left(\frac{\partial}{\partial \theta} \log f(x_j, \theta)\right) \right\} \\ &= n \cdot E\left[\frac{\partial}{\partial \theta} \log f(x, \theta)\right]^2, \quad [\text{On using } (*)] \end{aligned}$$

since  $x_i$ 's;  $i = 1, 2, \dots, n$  are i.i.d. r.v.'s.

### 15.7.1. Conditions for the Equality Sign in Cramer-Rao (C.R.) Inequality.

In proving (15.32) we used [c.f. (15.36) that

$$[\gamma'(\theta)]^2 \leq E[\iota - \gamma(\theta)]^2 \cdot E\left(\frac{\partial}{\partial \theta} \log L\right)^2 \quad \dots(15.39)$$

The sign of equality will hold in C.R. Inequality if and only if the sign of equality holds in (15.39). The sign of equality will hold in (15.39) by Cauchy Schwartz Inequality, if and only if the variables  $[\iota - \gamma(\theta)]$  and  $\left(\frac{\partial}{\partial \theta} \log L\right)$  are proportional to each other, i.e.,

$$\frac{\iota - \gamma(\theta)}{\frac{\partial}{\partial \theta} \log L} = \lambda = \lambda(\theta)$$

where  $\lambda$  is a constant independent of  $(x_1, x_2, \dots, x_n)$  but may depend on  $\theta$ .

$$\therefore \frac{\partial}{\partial \theta} \log L = \frac{t - \gamma(\theta)}{\lambda(\theta)} = [t - \gamma(\theta)] A(\theta) \quad \dots(15-40)$$

where  $A = A(\theta) = 1/\lambda(\theta)$ , say.

Hence a necessary and sufficient condition for an unbiased estimator  $t$  to attain the lower bound of its variance is given by (15-40).

Further, the C-R minimum variance bound is given by :

$$\text{Var}(t) = [\gamma'(\theta)]^2 / E \left( \frac{\partial}{\partial \theta} \log L \right)^2 \quad \dots(15-41)$$

$$\begin{aligned} \text{But } E \left( \frac{\partial}{\partial \theta} \log L \right)^2 &= E [A(\theta) \cdot \{ t - \gamma(\theta) \}]^2 \quad [\text{From (15-40)}] \\ &= [A(\theta)]^2 : E [t - \gamma(\theta)]^2 \\ &= [A(\theta)]^2 \cdot \text{Var}(t) \end{aligned}$$

Substituting in (15-41), we get

$$\begin{aligned} \text{Var}(t) &= \frac{[\gamma'(\theta)]^2}{[A(\theta)]^2 \cdot \text{Var}(t)} \\ \Rightarrow \text{Var}(t) &= \left| \frac{\gamma'(\theta)}{A(\theta)} \right| = |\gamma'(\theta) \lambda(\theta)| \quad \dots(15-42) \end{aligned}$$

Hence if the likelihood function  $L$  is expressible in the form (15-40) viz.,

$$\frac{\partial}{\partial \theta} \log L = \frac{t - \gamma(\theta)}{\lambda(\theta)} = [t - \gamma(\theta)] \cdot A(\theta),$$

then

- (i)  $t$  is an unbiased estimator of  $\gamma(\theta)$ ,
- (ii) Minimum Variance Bound (MVB) estimator ( $t$ ) for  $\gamma(\theta)$  exists, and
- (iii)  $\text{Var}(t) = \left| \frac{\gamma'(\theta)}{A(\theta)} \right| = |\gamma'(\theta) \lambda(\theta)|$

The importance of this result lies in that fact that C.R. inequality, in addition to find if MVB estimator for  $\gamma(\theta)$  exists, also gives us the variance of such an estimator, which is given by (15-42).

**Remarks 1.** If  $\gamma(\theta) = \theta$ , i.e., if  $t$  is an unbiased estimator of  $\theta$ , then (15-40) can be written as :

$$\frac{\partial}{\partial \theta} \log L = \frac{t - \theta}{\lambda} \quad \dots(15-43)$$

Hence if (15-43) holds, then  $t$  is an MVB estimator for  $\theta$  with

$$\text{Var}(t) = |\lambda(\theta)| = 1 / |A(\theta)| \quad \dots(15-43a)$$

**2.** We have seen in (15-40) that an MVB estimator exists for  $\gamma(\theta)$  if

$$\frac{\partial}{\partial \theta} \log L = \frac{t - \gamma(\theta)}{\lambda} = [t - \gamma(\theta)] \cdot \frac{1}{\lambda}, \quad \dots(*)$$

where  $\lambda = \lambda(\theta)$ ; say. If we write

$$\int \frac{1}{\lambda} d\theta = \alpha(\theta),$$

then integrating (\*) w.r. to  $\theta$  (by parts), we get

$$\begin{aligned}\log L &= [t - \gamma(\theta)] \alpha(\theta) + \int \alpha(\theta) \cdot \gamma'(\theta) d\theta + k(x) \\ \Rightarrow \log L &= [t - \gamma(\theta)] \alpha(\theta) + \beta(\theta) + k(x)\end{aligned} \quad \dots(15.44)$$

where  $\alpha(\theta)$  and  $\beta(\theta)$  are arbitrary functions of  $\theta$  and  $k(x) = k(x_1, x_2, \dots, x_n)$ , is an arbitrary function of  $x_i$ 's independent of  $\theta$ .

$$\begin{aligned}\text{Hence } \log f(x, \theta) &= [t - \gamma(\theta)] A_1(\theta) + B_1(\theta) + k_1(x) \\ \Rightarrow f(x, \theta) &= g(x) \cdot h(\theta) \cdot \exp [\alpha(\theta) \cdot \psi(x)]\end{aligned} \quad \dots(15.44a)$$

which is the necessary and sufficient condition for the existence of a sufficient statistic [c.f. Koopman's form, Equation (15.30) in Remark 3 to § 15.6]. Hence an MVB estimator for  $\gamma(\theta)$  exists if and only if there exists a sufficient estimator for  $\gamma(\theta)$ .

This suggests that in our search for an MVB estimator for  $\gamma(\theta)$ , we need to confine ourselves to sufficient estimators of  $\gamma(\theta)$  alone.

This explains why the method failed in the case of Cauchy population [c.f. Example 15.19], where no sufficient estimator exists and its success in the case of normal population [c.f. Example 15.18, where  $\bar{x}$  is sufficient for  $\mu$  and Example 15.20,  $\sum_{i=1}^n x_i^2/n$  is sufficient for  $\sigma^2$ ].

**Example 15.18.** Obtain the MVB estimator for  $\mu$  in the normal population  $N(\mu, \sigma^2)$ , where  $\sigma^2$  is known.

**Solution.** If  $x_1, x_2, \dots, x_n$  is a random sample of size  $n$  from the normal population, then

$$\begin{aligned}L &= \prod_{i=1}^n f(x_i, \mu) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \cdot \exp \left\{ - \sum_{i=1}^n (x_i - \mu)^2 / 2\sigma^2 \right\} \\ \log L &= -n \log (\sqrt{2\pi} \sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \\ &= k - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2,\end{aligned}$$

where  $k$  is a constant independent of  $\mu$ , ( $\sigma$  being known).

$$\begin{aligned}\frac{\partial}{\partial \mu} \log L &= -\frac{1}{2\sigma^2} \sum_{i=1}^n [2(x_i - \mu)(-1)] \\ &= \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} = \frac{\sum x_i - n\mu}{\sigma^2} = \frac{(\bar{x} - \mu)}{\sigma^2/n}\end{aligned}$$

which is of the form (15.40).

Hence  $\bar{x}$  is an MVB unbiased estimator for  $\mu$  and  $V(\hat{\mu}) = V(\bar{x}) = \frac{\sigma^2}{n}$ .

**Example 15-19.** Find if MVB estimator exists for  $\theta$  in the Cauchy's population :

$$dF(x, \theta) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}, -\infty < x < \infty,$$

**Solution.** Here

$$L = \prod_{i=1}^n f(x_i, \theta) = \left( \frac{1}{\pi} \right)^n \prod_{i=1}^n \left[ \frac{1}{1 + (x_i - \theta)^2} \right]$$

$$\therefore \log L = -n \log \pi - \sum_{i=1}^n \log [1 + (x_i - \theta)^2]$$

$$\Rightarrow \frac{\partial}{\partial \theta} \log L = 2 \sum_{i=1}^n \left[ \frac{(x_i - \theta)}{1 + (x_i - \theta)^2} \right]$$

Since this cannot be expressed in form (15-40), MVB estimator does not exist for  $\theta$ , in the Cauchy's population and so Cramer Rao lower bound is not attainable by the variance of any unbiased estimator  $\theta$ .

**Example 15-20.** A random sample  $x_1, x_2, \dots, x_n$  is taken from a normal population with mean zero and variance  $\sigma^2$ . Examine if  $\sum_{i=1}^n x_i^2/n$  is an MVB estimator for  $\sigma^2$ .

**Solution.** Since  $X \sim N(0, \sigma^2)$ ,

$$f(x, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \cdot \exp \left( -\frac{x^2}{2\sigma^2} \right), -\infty < x < \infty$$

$$L = \prod_{i=1}^n f(x_i, \sigma^2) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left\{ -\sum_{i=1}^n (x_i^2 / 2\sigma^2) \right\}$$

$$\Rightarrow \log L = -\frac{n}{2} \log (2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2$$

$$\frac{\partial}{\partial \sigma^2} \log L = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n x_i^2 = \frac{\sum_{i=1}^n x_i^2 - n\sigma^2}{2\sigma^4}$$

$$= \frac{\left( \sum_{i=1}^n x_i^2 / n \right) - \sigma^2}{(2\sigma^4 / n)},$$

which is of the form (15-40).

Hence  $\hat{\sigma}^2 = \sum_{i=1}^n \frac{x_i^2}{n}$ , is an MVB estimator and  $V(\hat{\sigma}^2) = \frac{2\sigma^4}{n}$

**Example 15.21.** Show that  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i / n$ , in random sampling from

$$f(x, \theta) = \begin{cases} (1/\theta) \exp(-x/\theta), & 0 < x < \infty \\ 0, & \text{otherwise} \end{cases} \quad \dots(*)$$

where  $0 < \theta < \infty$ , is an MVB estimator of  $\theta$  and has variance  $\theta^2/n$ .

**Solution.** Let  $x_1, x_2, \dots, x_n$  be a random sample of size  $n$  from population with p.d.f. in (\*). Then

$$L = \prod_{i=1}^n f(x_i, \theta) = \frac{1}{\theta^n} \cdot \exp \left[ - \sum_{i=1}^n x_i / \theta \right]$$

$$\Rightarrow \log L = -n \log \theta - \frac{1}{\theta} \cdot \sum_{i=1}^n x_i$$

$$\therefore \frac{\partial}{\partial \theta} \log L = -\frac{n}{\theta} + \frac{1}{\theta^2} \cdot \sum_{i=1}^n x_i$$

$$\Rightarrow \frac{\partial^2}{\partial \theta^2} \log L = \frac{n}{\theta^2} - \frac{2}{\theta^3} \cdot \sum_{i=1}^n x_i$$

$$\therefore I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \log L \right] = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \cdot \sum_{i=1}^n E(x_i)$$

In sampling from exponential population (\*), we have

$$E(X) = \theta \text{ and } \text{Var}(X) = \theta^2 \quad \dots(**)$$

$$\therefore I(\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \cdot \sum_{i=1}^n (\theta) \quad (\because x_i \text{'s are i.i. d.})$$

$$= -\frac{n}{\theta^2} + \frac{2}{\theta^3} \cdot n\theta = \frac{n}{\theta^2}$$

$$\text{Also } \gamma(\theta) = \theta \Rightarrow \gamma'(\theta) = 1.$$

Hence Cramer Rao lower bound to the variance of an unbiased estimator of  $\theta$  is :

$$\frac{[\gamma'(\theta)]^2}{I(\theta)} = \frac{1}{(n/\theta^2)} = \frac{\theta^2}{n} \quad \dots(***)$$

Consider the estimator  $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$ .

$$\text{We have : } E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(x_i) = \frac{1}{n} \sum_{i=1}^n (\theta) = \theta$$

$\Rightarrow \bar{X}$  is an unbiased estimator of  $\theta$ .

$$\text{Also } \text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{\text{Var } X}{n} = \frac{\theta^2}{n} \quad [\text{From } (**)]$$

Thus we see that  $\text{Var}(\bar{X})$  coincides with the Cramer-Rao lower bound obtained in (\*\*\*)). Hence  $\bar{X}$ , the sample mean is an MVB unbiased estimator, for  $\theta$ .

**Aliter.** A more convenient way of doing this problem is as follows :

We have

$$\begin{aligned}\frac{\partial}{\partial \theta} \log L &= -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n x_i = \frac{\sum_{i=1}^n x_i - n\theta}{\theta^2} \\ &= \frac{\bar{X} - \theta}{(\theta^2/n)} = \frac{\bar{X} - \theta}{\lambda(\theta)}, \text{ (say)}\end{aligned}$$

which is of the form (15-40).

Hence  $\bar{X}$  is an MVB unbiased estimator of  $\theta$  and  $\text{Var}(\bar{X}) = \lambda(\theta) = \theta^2/n$ .

**Example 15-22.** Given the probability density function

$$f(x : \theta) = [\pi \{1 + (x - \theta)^2\}]^{-1}; -\infty < x < \infty, -\infty < \theta < \infty \quad \dots (*)$$

show that the Cramer-Rao lower bound of variance of an unbiased estimator of  $\theta$  is  $\frac{2}{n}$ , where  $n$  is the size of the random sample from this distribution.

[Sri Venkateswara Univ M.Sc., 1992]

**Solution.**  $\log f = -\log \pi - \log [1 + (x - \theta)^2]$

$$\begin{aligned}\frac{\partial \log f}{\partial \theta} &= \frac{2(x - \theta)}{[1 + (x - \theta)^2]} \\ E\left(\frac{\partial \log f}{\partial \theta}\right)^2 &= \int_{-\infty}^{\infty} \frac{4(x - \theta)^2}{[1 + (x - \theta)^2]^2} f(x, \theta) dx \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{4(x - \theta)^2}{[1 + (x - \theta)^2]^3} dx\end{aligned}$$

Put  $x - \theta = \tan \phi \Rightarrow dx = \sec^2 \phi d\phi$ .

$$\begin{aligned}E\left(\frac{\partial \log f}{\partial \theta}\right)^2 &= \frac{2}{\pi} \int_0^{\pi/2} \frac{4 \tan^2 \phi}{\sec^6 \phi} \sec^2 \phi d\phi = \frac{2}{\pi} \int_0^{\pi/2} \frac{4 \sin^2 \phi}{\cos^4 \phi} \cos^4 \phi d\phi \\ &= \frac{2}{\pi} \int_0^{\pi/2} 4 \sin^2 \phi \cos^2 \phi d\phi = \frac{8}{\pi} \int_0^{\pi/2} (\cos^2 \phi - \cos^4 \phi) d\phi \\ &= \frac{8}{\pi} \left[ \frac{1}{2} \cdot \frac{\pi}{2} - \frac{3 \cdot 1}{4 \cdot 2} \cdot \frac{\pi}{2} \right]\end{aligned}$$

( Using reduction formula for  $\int_0^{\pi/2} \cos^n x dx$  ).

$$= \frac{8}{\pi} \left[ \frac{\pi}{4} - \frac{3\pi}{16} \right] = \frac{1}{2}$$

Hence Cramer-Rao lower bound is

$$= \frac{1}{n E \left( \frac{\partial \log f}{\partial \theta} \right)^2} = \frac{1}{n \left[ \frac{1}{2} \right]} = \frac{2}{n}.$$

**Examp<sup>l</sup>. 15.23.** Prove that under certain general conditions of regularity to be stated clearly the mean square deviation  $E(\hat{\theta} - \theta)^2$  of an estimator  $\hat{\theta}$  of the parameter  $\theta$ , can never fall below a positive limit depending only on the density function  $f(x, \theta)$ , the size of the sample and the bias of the estimate.

**Solution.** We have proved Cramer-Rao's inequality

$$V(\hat{\theta}) \geq \frac{[\psi'(\theta)]^2}{I(\theta)}, \text{ where } E(\hat{\theta}) = \psi(\theta). \quad \dots (*)$$

Now

$$\begin{aligned} E(\hat{\theta} - \theta)^2 &= E[\hat{\theta} - \psi(\theta) + \psi(\theta) - \theta]^2 \\ &= E[\hat{\theta} - \psi(\theta)]^2 + [\theta - \psi(\theta)]^2 + 2[\psi(\theta) - \theta] \cdot E[\hat{\theta} - \psi(\theta)] \\ &= V(\hat{\theta}) + [\theta - \psi(\theta)]^2 \\ \therefore E(\hat{\theta} - \theta)^2 &\geq \frac{[\psi'(\theta)]^2}{I(\theta)} + [\theta - \psi(\theta)]^2 \quad [\text{Using } (*)] \quad \dots (**). \end{aligned}$$

Let  $\hat{\theta}$  be a 'biased' estimator of  $\theta$  with bias given by  $b(\theta)$

$$\text{i.e., } E(\hat{\theta}) = \theta + b(\theta) = \psi(\theta), \text{ (say).}$$

$$\therefore \psi(\theta) - \theta = b(\theta)$$

From (\*\*), we get

$$\therefore E(\hat{\theta} - \theta)^2 \geq \frac{\left[ 1 + \frac{\partial}{\partial \theta} b(\theta) \right]^2}{I(\theta)} + [b(\theta)]^2 > 0,$$

$$\text{where } I(\theta) = n \int_{-\infty}^{\infty} \left( \frac{\partial}{\partial \theta} \log f \right)^2 f(x, \theta) dx > 0$$

This proves the result.

**15.8. Complete Family of Distributions.** Consider a statistic  $T = (x_1, x_2, \dots, x_n)$ , based on a random sample of size  $n$  from the population  $f(x, \theta)$ ,  $\theta \in \Theta$ . The distribution of the statistic  $T$  will, in general, depend on  $\theta$ . Hence corresponding to  $T$ , we again have a family of distributions, say,  $\{g(t, \theta), \theta \in \Theta\}$ .

**Definition.** The statistic  $T = t(x)$ , or more precisely the family of distributions  $\{g(t, \theta), \theta \in \Theta\}$  is said to be complete for  $\theta$  if

$$E_{\theta} [h(T)] = 0 \text{ for all } \theta \Rightarrow P_{\theta} [h(T) = 0] = 1 \quad \dots (15.45)$$

$$\left. \begin{array}{l} \text{i.e., } \int h(t) g(t, \theta) dt = 0 \text{ for all } \theta \in \Theta \\ \text{or } \sum_t h(t) g(t, \theta) = 0 \text{ for all } \theta \in \Theta \end{array} \right\} \quad \dots(15.45a)$$

$$\Rightarrow h(T) = 0, \text{ for all } \theta \in \Theta, \text{ almost surely (a.s.)}. \quad \dots(15.45b)$$

The concept of complete sufficient statistic is specially useful in Rao-Blackwell Theorem [c.f. § 15-9].

**Example 15-24.** Let  $X_1, X_2, \dots, X_n$  be a random sample from Bernoulli distribution :

$$f(x, \theta) = \begin{cases} \theta^x (1-\theta)^{1-x}; & x = 0, 1 \\ 0 & , \text{ otherwise} \end{cases}$$

Show that  $\sum_{i=1}^n X_i$ , is a complete sufficient statistic for  $\theta$ .

**Solution.** The likelihood function of the sample  $(X_1, X_2, \dots, X_n)$  is given by :  $L = \prod_{i=1}^n f(x_i, \theta) = \left[ \theta^{\sum x_i} (1-\theta)^{n - \sum x_i} \right] \times 1$

$$= g [t(x), \theta] \cdot h(x_1, x_2, \dots, x_n)$$

$$\text{where } t(x) = \sum_{i=1}^n x_i \quad \text{and} \quad h(x_1, x_2, \dots, x_n) = 1$$

Hence by Factorisation Theorem,  $T = \sum_{i=1}^n X_i$ , is sufficient estimator of  $\theta$ .

Since  $X_i$ 's are i.i.d. Bernoulli variates with parameter  $\theta$ ,

$$T = \sum_{i=1}^n X_i \sim B(n, \theta),$$

with p.m.f.

$$P(T=k) = \begin{cases} {}^n C_k \theta^k (1-\theta)^{n-k}, & k = 0, 1, 2, \dots, n \\ 0 & , \text{ otherwise} \end{cases}$$

$$E_\theta [h(T)] = \sum_{k=0}^n h(k) \cdot P(T=k) = \sum_{k=0}^n h(k) \cdot {}^n C_k \theta^k (1-\theta)^{n-k}$$

$$= \sum_{k=0}^n A(k) \cdot \theta^k (1-\theta)^{n-k}; \quad A(k) = h(k) \cdot {}^n C_k \quad \dots(*)$$

$$= A(0) (1-\theta)^n + A(1) \theta (1-\theta)^{n-1} + \dots + A(n) \theta^n$$

Now

$$E_\theta [h(T)] = 0 \text{ for all } \theta \in \Theta = \{\theta : 0 < \theta < 1\}$$

$$\Rightarrow A(0) (1-\theta)^n + A(1) \theta (1-\theta)^{n-1} + \dots + A(n) \theta^n = 0, \forall \theta$$

$$\Rightarrow A(0) + A_1 [\theta/(1-\theta)] + \dots + A(n) [\theta/(1-\theta)]^n = 0 \quad \forall \theta \in [0, 1] \quad \dots(**)$$

$$\Rightarrow A(0) = A(1) = A(2) = \dots = A(n) = 0,$$

since a polynomial of degree  $n$  in  $x$  is identically zero (for all  $x$ ), if all the coefficients are zero.

From (\*) and (\*\*), we get

$$h(k) = 0, k = 0, 1, 2, \dots, n$$

$$\Rightarrow h(t) = 0, t = 0, 1, 2, \dots, n$$

Hence  $T$  is a complete (sufficient) statistic for  $\theta$ .

**Example 15-25.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from  $N(\theta, 1)$  population. Examine if  $T = t(x) = X_1$  is complete for  $\theta$ .

**Solution.** We have  $T = X_1 ; \Theta = \{\theta : -\infty < \theta < \infty\}$

$$\therefore E_{\theta} [h(T)] = 0$$

$$\Rightarrow \int_{-\infty}^{\infty} h(u) e^{-(u-\theta)^2/2} du = 0, \text{ for all } \theta \in \Theta$$

$$\Rightarrow \int_{-\infty}^{\infty} \{h(u) e^{-u^2/2}\} e^{\theta u} du = 0, \text{ for all } \theta \in \Theta$$

This is a bilateral Laplace transform in  $\theta$ . Since these are unique :

$$h(u) \cdot e^{-u^2/2} = 0, \text{ a.s.}$$

$$\Rightarrow h(u) = 0, \text{ a.s.}$$

$$\Rightarrow P [h(T) = 0] = 1, \forall \theta \in \Theta$$

$$\Rightarrow T = X_1, \text{ is complete statistic for } \theta.$$

**Remark.** It can be easily seen that  $T_1 = \sum_{i=1}^n X_i$ , is a sufficient estimator of  $\theta$  and since  $T_1 \sim N(n\theta, 1/n)$ , by proceeding as in the above problem, we can prove that  $T_1 = \sum_{i=1}^n X_i$ , is a complete sufficient statistic for  $\theta$  and the family of distributions  $\{g_1(t_1, \theta), \theta \in \Theta\}$ , is complete.

**Example 15-26.** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(0, \theta)$ . Prove that  $T = X_1$  is not a complete statistic for  $\theta$  but  $T_1 = X_1^2$  is complete for  $\theta$ .

**Solution.** Here  $T = t(x) = X_1 ; \Theta = \{\theta ; 0 < \theta < \infty\}$

$$E_{\theta} [h(T)] = 0, \text{ for all } \theta \in \Theta$$

$$\Rightarrow \int_{-\infty}^{\infty} h(u) \exp [-u^2/(2\theta)] du = 0, \text{ for all } \theta \in \Theta$$

This holds only for all odd functions  $h(u)$  of  $u$ , for which the integral exists i.e., for all functions s.t.

- $h(u) = -h(-u)$ ; for all  $u$   
 $\Rightarrow h(u) \neq 0$ , a.s.  
 $\Rightarrow T = X_1$  is not complete statistic for  $\theta$ .

Let us now consider the statistic  $T_1 = X_1^2$ .

$$E_{\theta} [h(T_1)] = 0, \text{ for all } \theta \in \Theta$$

- $\Rightarrow \int_{-\infty}^{\infty} h(x^2) \exp(-x^2/2\theta) dx = 0, \text{ for all } \theta \in \Theta$   
 $\Rightarrow \int_{-\infty}^{\infty} \frac{h(u)}{\sqrt{u}} \exp(-u/2\theta) du = 0, \forall \theta \in \Theta$

This being a Laplace transform in  $(1/\theta)$ , we have

- $\frac{h(u)}{\sqrt{u}} = 0, \text{ a.s.}$   
 $\Rightarrow h(u) = 0, \text{ a.s.}$   
 $\Rightarrow T_1 = X_1^2$ , is complete statistic for  $\theta$ .

**Remark.** We can easily see that  $T_1 = X_1^2$ , is sufficient statistic for  $\theta$ . Hence  $T_1 = X_1^2$  is a complete sufficient statistic for  $\theta$ .

**Example 15-27.** Let  $X_1, X_2, \dots, X_n$  be a random sample from uniform  $U[0, \theta]$ ,  $\theta > 0$  population. Show that  $T = \max_{1 \leq i \leq n} (X_i) = X_{(n)}$ , is a complete sufficient statistic for  $\theta$ .

**Solution.**  $T = X_{(n)}$  has p.d.f.

$$g(t, \theta) = \begin{cases} \frac{n t^{n-1}}{\theta^n}; & 0 \leq t \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

$$E_{\theta} [h(T)] = 0, \text{ for all } \theta \in \Theta = \{\theta : 0 < \theta < \infty\}$$

$$\Rightarrow \frac{n}{\theta^n} \int_0^{\theta} h(u) \cdot u^{n-1} du = 0, \text{ for all } \theta \in \Theta$$

Differentiating w.r. to  $\theta$ , we get from the fundamental theorem of integral calculus :

- $h(\theta) \cdot \theta^{n-1} = 0, \forall \theta \in \Theta$   
 $\Rightarrow h(T) = 0, \text{ a.s.}$   
 $\Rightarrow T = \max (X_1, X_2, \dots, X_n) = X_{(n)}$ , is complete for  $\theta$ .

We have also proved in Example 15-13, that  $T = X_{(n)}$ , is sufficient for  $\theta$ . Hence  $T = X_{(n)}$ , is complete sufficient statistic for  $\theta$ .

**15-9. MVU and Blackwellisation.** Cramer-Rao inequality (c.f. § 15-7) provides us a technique of finding if the unbiased estimator is also an MVU estimator or not. Here, since the regularity conditions are very strict, its applications become quite restrictive. More-over MVU estimator is not the same as an MVU estimator since the Cramer-Rao lower bound may not always

be attained. More-over, if the regularity conditions are violated, then the least attainable variance may be less than the Cramer-Rao bound. [For illustration see Example 15-30]. In this section we shall discuss how to obtain MVU estimator from any unbiased estimator through the use of sufficient statistic. This technique is called Blackwellisation after D. Blackwell. The result is contained in the following Theorem due to C.R. Rao and D. Blackwell.

**Theorem 15-9. (Rao-Blackwell Theorem):** Let  $X$  and  $Y$  be random variables such that

$$E(Y) = \mu \text{ and } \operatorname{Var}(Y) = \sigma_Y^2 > 0$$

Let  $E(Y | X = x) = \phi(x)$ , then

$$(i) \quad E[\phi(X)] = \mu$$

$$\text{and (ii)} \quad \operatorname{Var}[\phi(X)] \leq \operatorname{Var}(Y)$$

**Proof.** Let  $f_{XY}(x, y)$  be the joint p.d.f. of random variables  $X$  and  $Y$ ,  $f_1(\cdot)$  and  $f_2(\cdot)$  the marginal p.d.f.'s of  $X$  and  $Y$  respectively and  $h(y | x)$  be the conditional p.d.f. of  $Y$  for given  $X = x$  such that

$$h(y | x) = \frac{f(x, y)}{f_1(x)}$$

$$\begin{aligned} E(Y | X = x) &= \int_{-\infty}^{\infty} y \cdot h(y | x) dy \\ &= \int_{-\infty}^{\infty} y \cdot \frac{f(x, y)}{f_1(x)} dy \\ &= \frac{1}{f_1(x)} \int_{-\infty}^{\infty} y f(x, y) dy = \phi(x), \text{ (say)} \end{aligned} \quad \dots(15-46)$$

$$\Rightarrow \int_{-\infty}^{\infty} y f(x, y) dy = \phi(x) \cdot f_1(x) \quad \dots(15-46a)$$

From (15-46) we observe that the conditional distribution of  $Y$  given  $X = x$  does not depend on the parameter  $\mu$ . Hence  $X$  is sufficient statistic for  $\mu$ .

Now

$$E[\phi(X)] = E[E(Y | X)] = E(Y) = \mu, \quad \dots(15-47)$$

which establishes part (i) of the Theorem.

We have

$$\begin{aligned} \operatorname{Var}(Y) &= E[Y - E(Y)]^2 = E[Y - \mu]^2 \\ &= E[Y - \phi(X) + \phi(X) - \mu]^2 \\ &= E[Y - \phi(X)]^2 + E[\phi(X) - \mu]^2 \\ &\quad + 2E[(Y - \phi(X)) (\phi(X) - \mu)] \end{aligned} \quad \dots(15-48)$$

The product term gives

$$\begin{aligned}
 E[(Y - \phi(X))(\phi(X) - \mu)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \phi(x))(\phi(x) - \mu) f(x, y) dx dy \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \phi(x))[\phi(x) - \mu] f_1(x) h(y|x) dx dy \\
 &= \int_{-\infty}^{\infty} [\phi(x) - \mu] \left[ \int_{-\infty}^{\infty} [y - \phi(x)] h(y|x) dy \right] dx
 \end{aligned}$$

But  $\int_{-\infty}^{\infty} [y - \phi(x)] h(y|x) dy = 0$        $\therefore E(Y|X=x) = \phi(x)$

$$\therefore E[(Y - \phi(X))(\phi(X) - \mu)] = 0$$

Substituting in (15-48), we get

$$\text{Var}(Y) = E[Y - \phi(X)]^2 + \text{Var}[\phi(X)] \quad \dots(15-49)$$

$$\Rightarrow \text{Var } Y \geq \text{Var}[\phi(X)] \quad (\because E[Y - \phi(X)]^2 \geq 0)$$

$$\Rightarrow \text{Var}[\phi(X)] \leq \text{Var } Y, \quad \dots(15-49a)$$

which completes the proof of the theorem.

**Remarks.** 1. From (15-49), it is obvious that the sign of equality holds in (15-49a) iff

$$\begin{aligned}
 E[Y - \phi(X)]^2 &= 0 \\
 \Rightarrow Y - \phi(X) &= 0, \text{ almost surely.} \\
 \text{i.e., iff} \quad P\{(x, y) : y - \phi(x) = 0\} &= 1 \quad \dots(15-50)
 \end{aligned}$$

2. Here we have proved the theorem for continuous r.v.'s. The result can be similarly proved for discrete case, replacing integration by summation.

3. Rao-Blackwell theorem enables us to obtain MVU estimators through sufficient statistic. If a sufficient estimator exists for a parameter, then in our search for MVU estimator we may restrict ourselves to functions of the sufficient statistic. The theorem can be stated slightly different as follows :

Let  $U = U(x_1, x_2, \dots, x_n)$  be an unbiased estimator of parameter  $\gamma(\theta)$  and let  $T = T(x_1, x_2, \dots, x_n)$  be a sufficient statistic for  $\gamma(\theta)$ . Consider the function  $\phi(T)$  of the sufficient statistic defined as

$$\phi(t) = E(U|T=t) \quad \dots(15-51)$$

which is independent of  $\theta$  (since  $T$  is sufficient for  $\gamma(\theta)$ ). Then

$$E\phi(T) = \gamma(\theta)$$

$$\text{and} \quad \text{Var } \phi(T) \leq \text{Var } (U) \quad \dots(15-52)$$

This result implies that starting with an unbiased estimator  $U$ , we can improve upon it by defining a function  $\phi(T)$  of the sufficient statistic as given in (15-51). This technique of obtaining improved estimators is called Blackwellisation.

If in addition, the sufficient statistic  $T$  is also complete, then the estimator  $\phi(T)$  discussed above will not only be an improved estimator over  $U$  but also the best (unique) estimator. We state below the relevant theorem.

**Theorem 15-10.** Let  $T$  be a complete sufficient statistic for  $\gamma(\theta)$ ,  $\theta \in \Theta$ . Then  $\phi(T)$ , the function of  $T$  defined in (15-51) is the unique unbiased estimator of  $\gamma(\theta)$ .

Combining the results of the two Theorems 15-9 and 15-10, we have the following result..

**Corollary.** If  $T$  is a complete sufficient statistic for  $\gamma(\theta)$  and if we can find some function of  $T$ , say  $g(T)$ , which is unbiased estimator of  $\gamma(\theta)$ , then  $g(T)$  is the MVU estimator of  $\gamma(\theta)$ .

**Example 15-28.** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\theta, 1)$ . Obtain MVUE of  $\theta$ .

**Solution.** it can be easily proved [c.f. Example 15-25] that the statistic

$$T = X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_i$$

is complete sufficient statistic for  $\theta$ .

$$\text{Consider } \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = \frac{T}{n} = g(T), \text{ (say)}$$

Since  $\bar{X}_n = g(T)$ , is unbiased estimator of  $\theta$ , by corollary to Theorem 15-10,  $\bar{X}_n$  is MVUE of  $\theta$ .

**Example 15-29.** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $U[0, \theta]$  population. Obtain MVUE for  $\theta$ .

**Solution.** We have seen that in sampling from  $U[0, \theta]$  population, the statistic :

$$T = X_{(n)} = \max_{1 \leq i \leq n} (X_i)$$

is sufficient (Example 15-13) and complete (Example 15-27) for  $\theta$ . Also

$$E(T) = E[X_{(n)}] = \left( \frac{n}{n+1} \right) \theta \quad [\text{See Example 15-30}]$$

$$\Rightarrow E\left[\frac{(n+1)T}{n}\right] = \theta$$

Hence by corollary to Theorem 15-10,  $[(n+1)T/n] = [(n+1)X_{(n)}/n]$  is an MVU estimator of  $\theta$ .

**Example 15-30.** Given :

$$\begin{aligned} f(x, \theta) &= \frac{1}{\theta}, \quad 0 < x < \theta, \quad \theta > 0 \\ &= 0, \text{ elsewhere,} \end{aligned} \quad \dots (*)$$

compute the reciprocal of

$$n E\left[\left(\frac{\partial \log f(x, \theta)}{\partial \theta}\right)^2\right]$$

and compare this with the variance of  $(n+1) Y_n/n$ , where  $Y_n$  is the largest item of a random sample of size  $n$  from this distribution. Comment on the result.

**Solution.**  $\log f(x, \theta) = -\log \theta \Rightarrow \frac{\partial}{\partial \theta} \log f = -\frac{1}{\theta}$

$$\Rightarrow n E\left(\frac{\partial}{\partial \theta} \log f\right)^2 = n E\left(\frac{1}{\theta^2}\right) = \frac{n}{\theta^2}$$

$$\text{Hence reciprocal of } n E\left[\left(\frac{\partial}{\partial \theta} \log f(x, \theta)\right)^2\right] = \frac{\theta^2}{n} \quad \dots (**)$$

For the rectangular population (\*), the p.d.f. of  $n$ th order statistic (the largest sample observation),  $Y_n$  is

$$g(y) = n \cdot [F(y, \theta)]^{n-1} f(y, \theta)$$

where  $F(x, \theta) = P(X \leq x) = \int_0^x f(u) du = \int_0^x \frac{1}{\theta} = \frac{x}{\theta}$

$$g(y) = n \left(\frac{y}{\theta}\right)^{n-1} \frac{1}{\theta} = \frac{n}{\theta^n} \cdot y^{n-1}; 0 \leq y < \theta$$

$$E(Y_n) = \int_0^\theta y^r \cdot g(y) dy = \frac{n}{\theta^n} \int_0^\theta y^{r+n-1} dy = \frac{n \theta^r}{n+r}$$

Taking  $r = 1$  and 2, we get

$$E(Y_n) = \frac{n\theta}{n+1}; E(Y_n^2) = \frac{n\theta^2}{n+2} \quad \dots (***)$$

Now  $E\left[\frac{n+1}{n} \cdot Y_n\right] = \frac{n+1}{n} E(Y_n) = \theta$  [Using \*\*\*]

$\Rightarrow (n+1)Y_n/n$  is an unbiased estimator of  $\theta$ .

$$\begin{aligned} \text{Var}\left[\frac{n+1}{n} Y_n\right] &= \left(\frac{n+1}{n}\right)^2 \cdot \text{Var}(Y_n) \\ &= \left(\frac{n+1}{n}\right)^2 [EY_n^2 - (EY_n)^2] \end{aligned}$$

$$= \left(\frac{n+1}{n}\right)^2 \left[ \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} \right] \quad \text{[Using (***)]}$$

$$= \theta^2 \left[ \frac{(n+1)^2}{n(n+2)} - 1 \right] = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n}$$

$$\Rightarrow \text{Var}\left[\frac{n+1}{n} \cdot Y_n\right] \leq 1 \sqrt{n E\left(\frac{\partial}{\partial \theta} \log f\right)^2}$$

Hence  $(n+1)Y_n/n$  is an MVUE.

**Remark.** This example illustrates that if the regularity conditions underlying Cramer-Rao inequality are violated, then the least attainable variance may be less than the Cramer-Rao lower bound.

### EXERCISE 15(a)

- What do you understand by Point Estimation? Define the following terms and give one example for each:

- (i) Consistent Statistic
- (ii) Unbiased Statistic
- (iii) Sufficient Statistic
- (iv) Efficiency.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1982*]

2. What do you understand by Point Estimation ? When would you say that estimate of a parameter is good ? In particular, discuss the requirements of consistency and unbiasedness of an estimate. Give an example to show that a consistent estimate need not be unbiased.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986*]

3. Discuss the terms (i) estimate, (ii) consistent estimate, (iii) unbiased estimate, of a parameter and show that sample mean is both consistent and unbiased estimate of the population mean.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1986*]

4. (a) If  $s_1^2, s_2^2, \dots, s_r^2$  are  $r$  sample variances based on random samples of sizes  $n_1, n_2, \dots, n_r$ , respectively, and if  $T$  is some statistic given by

$$T = \frac{n_1 s_1^2 + n_2 s_2^2 + \dots + n_r s_r^2}{a},$$

for estimating  $\sigma^2$  as an unbiased estimator, find the value of  $a$ , supposing population is very large and for every sample

$$s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$\text{Ans. } a = (n_1 + n_2 + \dots + n_r) - r.$$

(b) If  $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots, \bar{X}_r$  are the sample means based on samples of sizes  $n_1, n_2, n_3, \dots, n_r$ , respectively, an unbiased estimator,

$$t = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \dots + n_r \bar{X}_r}{k}$$

has been defined to estimate  $\mu$ . Find the value of  $k$ .

$$\text{Ans. } k = n_1 + n_2 + \dots + n_r.$$

5. (a) For the geometric distribution,

$$f(x, \theta) = \theta (1 - \theta)^{x-1}, \quad (x = 1, 2, \dots), \quad 0 < \theta < 1,$$

Obtain an unbiased estimator of  $1/\theta$ .

$$[\text{Ans. } E(\bar{X}) = 1/\theta.]$$

(b) The random variable  $X$  takes the values 1 and 0 with respective probabilities  $\theta$  and  $1 - \theta$ . Independent observations  $X_1, X_2, \dots, X_n$  on  $X$  are available. Write  $\xi = X_1 + X_2 + \dots + X_n$ .

Show that  $\xi(n - \xi)/n(n - 1)$  is an unbiased estimate of  $\theta(1 - \theta)$ .

6. Show that if  $T$  is an unbiased estimator of a parameter  $\theta$ , then  $\lambda_1 T + \lambda_2$  is an unbiased estimator of  $\lambda_1 \theta + \lambda_2$ , where  $\lambda_1$  and  $\lambda_2$  are known constants, but  $T^2$  is a biased estimator of  $\theta^2$ .

7. For the following cases determine if the given estimator is unbiased for the parametric function. When it is biased, derive an unbiased estimator from it.  $\bar{x}$  is the sample mean.

(a)  $x_1, \dots, x_n$  is a random sample from a distribution with variance  $\sigma^2$ . The estimator  $n^{-1} [(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2]$  is used to estimate  $\sigma^2$ .

(b)  $x_1, \dots, x_n$  is an independent sample from an exponential distribution with mean  $\theta$ . The estimator  $\left(1 - \frac{1}{n\bar{X}}\right)^{n-1}$  is used to estimate  $\exp\left(-\frac{1}{\theta}\right)$  when  $n\bar{X} > 1$  and zero is used when  $n\bar{X} < 1$ .

(c)  $r$  successes are observed in  $n$  Bernoulli trials with success probability  $p$ .  $(r/n)^2$  is used to estimate  $p^2$ .

$$8. f(x ; \mu, \sigma) = \frac{1}{\sigma} \exp\left[-\left(\frac{x - \mu}{\sigma}\right)\right]; \mu \leq x < \infty, -\infty < \mu < \infty \text{ and } 0 < \sigma < \infty$$

Obtain

- (i) an unbiased estimate of  $\mu$  when  $\sigma$  is known,
- (ii) an unbiased estimate of  $\sigma$  when  $\mu$  is known,
- (iii) two unbiased estimators of  $\sigma^2$  when  $\mu$  is known.

Hence obtain an infinity of unbiased estimators of  $\sigma^2$  in this case.

[Hint. The exponential distribution has mean  $\mu + \sigma$  and variance  $\sigma^2$ ]

9. Suppose  $X$  and  $Y$  are independent random variables with the same unknown means  $\mu$ . Both  $X$  and  $Y$  have variance as 36. Let  $T = aX + bY$  be an estimator of  $\mu$ .

(i) Show that  $T$  is an unbiased estimator of  $\mu$  if  $a + b = 1$ .

(ii) If  $a = \frac{1}{3}$  and  $b = \frac{2}{3}$ , what is the variance of  $T$ ?

(iii) If  $a = \frac{1}{2}$  and  $b = \frac{1}{2}$ , what is the variance of  $T$ ?

(iv) What choice of  $a$  and  $b$  minimizes the variance of  $T$  subject to the requirement that  $T$  is an unbiased estimate of  $\mu$ ?

10. (a) Examine the unbiasedness of the following estimates :

$$(i) s_1^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$(ii) s_2^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \text{ (where } \mu \text{ is known),}$$

for  $\sigma^2$ , the population variance.

[Delhi Univ. B.Sc. (Stat. Hons.), 1982]

$$\text{Ans. } E(s_1^2) = \left(\frac{n-1}{n}\right) \sigma^2 \neq \sigma^2, (ii) E(s_2^2) = \sigma^2.$$

(b) Let  $x_1, x_2, \dots, x_n$  be a random sample of size  $n$  drawn from a population with mean  $\mu$  and variance  $\sigma^2$ . Obtain an unbiased estimator for  $\mu^2$ .

**Hint.**  $E(\bar{X}) = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \mu ; \text{Var}(\bar{X}) = \sigma^2/n$

$$E(\bar{X}^2) = \text{Var}(\bar{X}) + [E(\bar{X})]^2 = \mu^2 + (\sigma^2/n)$$

**Ans.**  $\bar{X}^2 - (\sigma^2/n)$ , if  $\sigma^2$  is known;

and  $\bar{X}^2 - (S^2/n) = \bar{X}^2 - \frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$ , if  $\sigma^2$  is unknown.

11. If  $X_1, X_2, \dots, X_n$  is a random sample of size  $n$  from  $N(\mu, \sigma^2)$ , where  $\mu$  is known and if

$$T = \frac{1}{n} \sum_{i=1}^n |X_i - \mu|$$

examine if  $T$  is unbiased for  $\sigma$ . If not, obtain an unbiased estimator of  $\sigma$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

**Ans.** Hint.  $E(T) = \frac{1}{n} \sum_{i=1}^n E|X_i - \mu| = \sqrt{(2/\pi)} \cdot \sigma$

since, for  $N(\mu, \sigma^2)$ , Mean Deviation about mean =  $\sqrt{(2/\pi)} \sigma$

**Ans.** No;  $\sqrt{(\pi/2)} T$ .

12. If  $x_1, x_2, \dots, x_n$  is a random sample from the population

$$f(x, \theta) = (\theta + 1) x^\theta; 0 < x < 1; \theta > -1 \quad \dots (*)$$

show that  $\left[ \frac{-(n-1)}{\sum \log x_i} - 1 \right]$  is unbiased estimator of  $\theta$ .

**Hint.** In sampling from (\*),  $U = -\log X$  has an exponential distribution with parameter  $(\theta + 1)$

$$\Rightarrow U_i = -\log X_i \stackrel{i.i.d.}{\sim} \gamma(\theta + 1, 1); i = 1, 2, \dots, n$$

$$\Rightarrow Y = -\sum_{i=1}^n \log X_i \sim \gamma(\theta + 1, n); E[1/Y] = \frac{\theta + 1}{n-1}$$

13. Suppose  $X$  has a truncated Poisson distribution with p.m.f.

$$f(x, \theta) = \begin{cases} \frac{\exp(-\theta) \cdot \theta^x}{[1 - \exp(-\theta)] x!}, & x = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

Show that the only unbiased estimator of  $[1 - \exp(-\theta)]$  based on  $X$  is the statistic  $T$ , defined as :

$$T(x) = \begin{cases} 0, & \text{when } x \text{ is odd} \\ 2, & \text{when } x \text{ is even} \end{cases}$$

**Note.** This is an Example of absurd unbiased estimator.

14. Consider a random sample  $X_1, X_2, X_3$  of size 3 from uniform p.d.f.

$$f(x; \theta) = \begin{cases} 1/\theta, & 0 < x < \theta \\ 0, & \text{otherwise} \end{cases}$$

Show that each of the statistics  $4X_{(1)}$ ,  $2X_{(2)}$  and  $\frac{1}{3}X_{(3)}$ , where  $X_{(i)}$  is the  $i$ th order statistic is an unbiased estimator for  $\theta$ . Find the variance and hence the efficiency of each.

15. Obtain an unbiased estimator for (i)  $\theta$ , and (ii)  $\theta^2$ , in case of binomial probability distribution :

$$f(x; \theta) = {}^n C_x \theta^x (1-\theta)^{n-x}; x = 0, 1, 2, \dots, n; 0 < \theta < 1.$$

Hint.  $E\left(\frac{x}{n}\right) = \theta$ ;  $E\left[\frac{x(x-1)}{n(n-1)}\right] = \theta^2$ .

If we write  $T = x/n$ , the observed proportion of successes then

$$E(T) = \theta; E(T^2) = \frac{\theta^2}{n} + \left(\frac{n-1}{n}\right)\theta^2 \neq \theta^2.$$

This illustrates that we may have :

$t_n$  unbiased for  $\theta$  but  $t_n^2$  not unbiased for  $\theta^2$ .

16. Define 'efficiency of an estimator'.

$X$  is a uniform random variable with range  $[0, \theta]$ .  $x_1, x_2, \dots, x_n$  are independent observations on  $X$ . Define

$$\hat{\theta}_1 = \frac{2}{n}(x_1 + x_2 + \dots + x_n); \hat{\theta}_2 = \left[ \frac{(n+1)}{n} \right] \max(x_1, x_2, \dots, x_n).$$

Show that  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are unbiased for  $\theta$ . Evaluate their relative efficiency.

17. (a) The observations  $x_1, x_2, \dots, x_n$  represent a random sample from a uniform distribution over the interval  $(0, \theta)$ , where  $\theta$  is an unknown parameter.

The statistics  $\bar{X}$ ,  $m$  and  $M$  are the mean, the smallest value and the largest value respectively for the sample. Find values for  $k$  so that,  $kt$  is an unbiased estimator for  $\theta$  where

- (a)  $t = \bar{X}$
- (b)  $t = M$ ,
- (c)  $t = M - m$ ,

Of the three unbiased estimators which is the best? Give your reasons.

(b) Let  $X_1, X_2, \dots, X_n$  ( $n > 2$ ) be a random sample of size  $n$  from the distribution having density function :-

$$f(x; \theta) = \theta x^{\theta-1}, 0 < x < 1, \theta > 0$$

If  $Z = -\sum_{i=1}^n \log X_i$ , show that  $\frac{n-1}{Z}$  is an unbiased estimator for  $\theta$  and its efficiency is  $(n-2)/n$ .

Hint. See hint to Question 12.

18. (a) Suppose  $X_1, X_2, \dots, X_n$  are sample values independently drawn from population with mean  $m$  and variance  $\sigma^2$ . Consider the estimates :-

$$Y_n = \frac{X_1 + X_2 + \dots + X_n}{n+1}, \quad Z_n = \frac{X_1 + 2X_2 + 3X_3 + \dots + nX_n}{n^2}.$$

Discuss whether they are unbiased, consistent for  $m$ . What is the efficiency of  $Y_n$  over  $Z_n$ ?

(b) Let  $X_1, X_2, X_3$  and  $X_4$  be independent random variables such that  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2$  for  $i = 1, 2, 3, 4$ .

$$\text{If } Y = \frac{X_1 + X_2 + X_3 + X_4}{4}, \quad Z = \frac{X_1 + X_2 + X_3 + X_4}{5}$$

$$\text{and } T = \frac{X_1 + 2X_2 + X_3 - X_4}{4},$$

examine whether  $Y, Z$  and  $T$  are unbiased estimators of  $\mu$ ? What is the efficiency of  $Y$  relative to  $Z$ ?

(c) Let  $x_1, x_2, x_3, x_4$ , be a random sample from a  $N(\mu, \sigma^2)$  population. Find the efficiency of  $T = \frac{1}{7}(x_1 + 3x_2 + 2x_3 + x_4)$  relative to  $\bar{X} = \frac{1}{4} \sum_1^4 x_i$ . Which is relatively more efficient? Why?

19. A simple random sample of size 2 is drawn from a population containing 3 units, without replacement. Let  $y_1, y_2, y_3$  be the value of a characteristic measured on the three units and let  $T_{ij}$  be the estimator of the population mean  $\bar{Y}$  for the sample that has units  $i$  and  $j$ ;  $i, j = 1, 2, 3, i \neq j$ .

If  $T_{12} = (y_1 + y_2)/2$ ,  $T_{13} = (y_1/2) + (2y_3/3)$ ,  $T_{23} = (y_2/2) + (y_3/3)$ , show that  $T_{ij}$  is unbiased for  $\bar{Y}$ . Find the variance of  $T_{ij}$  and hence show that the variance of  $T_{ij}$  is smaller than that of the sample mean estimator if  $y_3(3y_2 - 3y_1 - y_3) = 0$ . [Indian Forest Service, 1991]

20. Let  $x$ , the earnings of a commercial bank, be a random variable with mean  $\mu$  and variance  $\sigma^2$ . A random sample of earnings of  $n$  banks is denoted by  $x_1, x_2, \dots, x_n$ . However, because of the disclosure laws, individual bank earnings are not disclosed and only the following average values are made available to the researcher :

$$a_1 = \frac{x_1 + x_2}{2}, \quad a_2 = \frac{x_3 + x_4}{2}, \quad \dots, \quad a_m = \frac{x_{n-1} + x_n}{2},$$

where  $n$  is an even number and  $m = n/2$ .

(i) Devise the best linear unbiased estimator of  $\mu$ , given the available information. What is the variance of the proposed estimator?

(ii) Devise an unbiased estimator of  $\sigma^2$ . [Delhi Univ. M.A. (Eco.), 1990]

21. (a) Define a consistent estimator.

Let  $T_n$  be an estimator of  $\theta$  with variance  $\sigma_n^2$  and  $E(T_n) = \theta_n$ . Prove that if  $\theta_n \rightarrow \theta$  and  $\sigma_n^2 \rightarrow 0$ , as  $n \rightarrow \infty$  then  $T_n$  is a consistent estimator of  $\theta$ .

Hence obtain consistent estimators for :

(i) Mean of the normal distribution.

(ii) Variance of the normal distribution when mean is known.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

(b) Give an example of an estimator :

(i) which is consistent but not unbiased,

(ii) which is unbiased but not consistent.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

22. (a) State and prove a sufficient condition for the consistency of an estimator. Define the invariance property of a consistent estimator and establish it.

[Delhi Univ. B.Sc. (Stat. Hons.), 1985]

(b) Given a random sample  $X_1, X_2, \dots, X_n$  from a normal  $(\mu, \sigma^2)$  distribution, examine unbiasedness and consistency of

$$(i) \bar{X} \text{ for } \mu, \quad (ii) \frac{1}{n} \sum (X_i - \bar{X})^2 \text{ for } \sigma^2.$$

23. (a) When would you say that estimate of a parameter is good ? In particular, discuss the requirements of consistency and unbiasedness of an estimate. Give an example to show that a consistent estimate need not be unbiased.

Show that an unbiased estimator whose variance tends to zero as the sample size increases to infinity is consistent.

(b) Define unbiasedness and consistency of estimators. Let  $X_1, X_2, \dots, X_n$  be a random sample from the  $N(\mu, \sigma^2)$  distribution. Propose three estimators of  $\mu$  based on this random sample such that the first is unbiased but not consistent, the second is consistent but not unbiased and the third is both unbiased and consistent.

[Punjab Univ. M.A. (Eco.), 1990]

24. (a) Define an unbiased and consistent estimate of a parameter in a population distribution.

Prove that for a sample of size  $n$  from a normal  $(m, 1)$  population, the arithmetic mean is an unbiased estimate of  $m$  and by Chebyshev's inequality or otherwise, show that the estimate is consistent too.

[Calcutta Univ. B.Sc. (Maths. Hons.), 1991]

(b) If  $X_1, X_2, \dots, X_n$  is a random sample obtained from the density function :

$$f(x, \theta) = 1, \quad \theta < x < \theta + 1 \\ = 0, \quad \text{elsewhere}$$

show that the sample mean  $\bar{X}$  is an unbiased and consistent estimator of  $\theta + \frac{1}{2}$ .

25. (a) Define a consistent estimator. Let  $T_{1,n}$  and  $T_{2,n}$  be consistent estimators of  $g_1(\theta)$  and  $g_2(\theta)$  respectively. Prove that  $aT_{1,n} + bT_{2,n}$  is a consistent estimator of  $ag_1(\theta) + bg_2(\theta)$ , where  $a$  and  $b$  are constants independent of  $\theta$ .

(b) Define consistent estimator. If the estimator  $t_n$  based on a random sample of size  $n$  is such that

$$E(t_n) \rightarrow \theta$$

and

$$V(t_n) \rightarrow 0,$$

as  $n \rightarrow \infty$ , then prove that  $t_n$  is a consistent estimator for  $\theta$ . Hence prove that sample mean is always a consistent estimate for population mean.

[Delhi Univ. M.Sc. (Maths), 1990]

(c) If  $t_n$  is a biased estimate of parameter  $\theta$  based on a random sample of size  $n$ , and  $E(t_n) = \theta + b_n$  and if  $b_n \rightarrow 0$  and  $V(t_n) \rightarrow 0$  as  $n \rightarrow \infty$ , show that  $t_n$  is consistent estimator of  $\theta$ .

(d) Define a consistent estimator of parameter  $\theta$ . If  $T$  is a consistent estimator of  $\theta$  and if  $\phi$  is any continuous function of its argument, show that  $\phi(T)$  is a consistent estimator of  $\phi(\theta)$ .

26. (a) Show that  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  and  $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ , are joint

consistent estimators for  $\mu$  and  $\sigma^2$  respectively, if  $x_1, x_2, \dots, x_n$  is a random sample from a normal population  $N(\mu, \sigma^2)$ .

Also find the efficiency of  $n s^2 / (n - 1)$ .

(b) Show that if  $t$  is a consistent estimator of a parameter  $\theta$ , then  $e^t$  is a consistent estimator of  $e^\theta$ .

(c) Prove that in case of Binomial distribution with parameter  $\theta$ ,  $t_n$  defined as  $r/n$  is a consistent unbiased estimator for  $\theta$ , but  $t_n$  defined as  $(r/n)^2$  is consistent but not unbiased estimator for  $\theta^2$ .

27. Show that in sampling from Cauchy distribution

$$f(x, \theta) = \frac{1}{\pi[1 + (x - \theta)^2]}, \quad -\infty < x < \infty, \quad \theta > 0;$$

(i) Sample mean  $\bar{X}$  is not a consistent estimator of  $\theta$ .

(ii) Sample median is a consistent estimator of  $\theta$  and its asymptotic efficiency is  $8/\pi^2$ .

28. (a) If  $T_1$  and  $T_2$  are consistent estimators of  $\gamma(\theta)$ , show that  $a_1 T_1 + a_2 T_2$ , such that  $a_1 + a_2 = 1$ , is also consistent for  $\gamma(\theta)$ .

(b) For a Poisson distribution with parameter  $\theta$ , show that  $1/\bar{X}$  is consistent estimator of  $1/\theta$ , where  $\bar{X}$  is the mean of a random sample from the given population.

**Hint.** Prove that  $\bar{X}$  is a consistent estimator of  $\theta$  and then use Invariance Property (Theorem 15-1).

29. Define MVU estimator. If  $T_1$  and  $T_2$  are two unbiased estimators of a parameter  $\theta$ , with variances  $\sigma_1^2$  and  $\sigma_2^2$  and correlation coefficient  $\rho$ , then obtain the best unbiased linear combination of  $T_1$  and  $T_2$ . Also obtain its variance. [Delhi Univ. B.Sc. (Stat. Hons.), 1990]

30. (a) Let  $T_1$  and  $T_2$  be two unbiased estimators of  $\gamma(\theta)$  having the same variance. Show that their correlation coefficient  $\rho_\theta$  cannot be smaller than  $(2e_\theta - 1)$ , where  $e_\theta$  is the efficiency of each estimator.

Further show that if  $T_1$  is MVU estimator and  $T_2$  is any unbiased estimator with efficiency  $e$ , then

$$V(T_1 - T_2) = \left( \frac{1}{e} - 1 \right) V(T_1)$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

(b) If  $T_1$  is a MVU for  $\theta$  and  $T_2$  is any other unbiased estimator of  $\theta$  with efficiency  $e_\theta$  then prove that the correlation between  $T_1$  and  $T_2$  is  $\sqrt{e_\theta}$ .

[Delhi Univ. B.A. (Stat. Hons.), 1987]

31. (a) Define MVU estimator. Show that an MVU estimator is unique.

[Delhi Univ. B.Sc. (Stat. Hons.), 1985]

(b) If  $T_1$  and  $T_2$  are two unbiased statistics having the same variance and  $\rho$  is the correlation between them then show that  $\rho \geq 2e - 1$ , where  $e$  is the ratio of the variance of the best estimator to the common variance of  $T_1$  and  $T_2$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1992]

32. (a) Let  $T$  be an MVU estimate for  $\gamma(\theta)$  and  $T_1, T_2$  be two other unbiased estimators of  $\gamma(\theta)$  with efficiencies  $e_1$  and  $e_2$  respectively.

If  $\rho_\theta$  is the correlation coefficient between  $T_1$  and  $T_2$ , then

$$(e_1 e_2)^{1/2} - ((1 - e_1)(1 - e_2))^{1/2} \leq \rho_\theta \leq (e_1 e_2)^{1/2} + ((1 - e_1)(1 - e_2))^{1/2}.$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1988, 1986]

(b) Let  $t_1$  and  $t_2$  be two unbiased estimates of  $\theta$  with variances  $\sigma_1^2$  and  $\sigma_2^2$ , (both known) and correlation  $\rho$  (known). Consider the estimate

$$\hat{\theta} = \alpha t_1 + (1 - \alpha) t_2.$$

Show that  $\hat{\theta}$  is unbiased. Find  $\alpha$  such that  $\hat{\theta}$  has minimum variance.

[Delhi Univ. M.A. (Eco.), 1986]

33. Suppose  $X$  and  $Y$  are independent unbiased estimates of  $\mu$ . It is known that the variance of  $X$  is 12 and the variance of  $Y$  is 4. It is desired to combine two estimators in order to obtain a more efficient estimator: Let  $T = aX + bY$ , be the new estimator.

(i) In order that  $T$  be an unbiased estimator of  $\mu$ , what conditions must be imposed on  $a$  and  $b$ ?

(ii) Find the values of  $a$  and  $b$  that minimize the variance of  $T$  subject to the condition that  $T$  be an unbiased estimator.

34. (a) What is an efficient estimator ?

If  $T_1, T_2$  are both efficient estimators with variance  $v$  and if  $T = \frac{1}{2}(T_1 + T_2)$ , show that variance of  $T$  is  $(v/2)(1 + \rho)$ , where  $\rho$  is the coefficient of correlation between  $T_1$  and  $T_2$ . Deduce that  $\rho = 1$  and that  $T$  is also efficient.

(b) If  $T$  and  $T'$  be two consistent estimators of which  $T$  is the most efficient, prove that the correlation coefficient between them is

$$\sqrt{\frac{V(T)}{V(T')}} , \text{ where } V(T) \text{ and } V(T') \text{ are the variance of } T \text{ and } T' \text{ respectively.}$$

Show also that the correlation coefficient between two most efficient estimators is unity.

[Allahabad Univ. M.A. (Eco.), 1993]

35. Define a sufficient statistic. Explain the method of finding sufficient estimator. If  $(x_1, x_2, \dots, x_n)$  is a random sample from a distribution :

$$f(x, p) = p^x (1-p)^{1-x}; x = 0, 1 \text{ and } 0 \leq p \leq 1,$$

find the sufficient estimator of  $p$ .

[Madras Univ. B.Sc., 1988]

36. State the factorisation theorem on sufficiency. Obtain a sufficient statistic for the parameter  $\theta$  in the following distribution :

$$f(x : \theta) = \frac{1}{\theta}, \quad 0 < x < \theta.$$

(b) Define a sufficient statistic.

If  $x_1, x_2, \dots, x_n$  is a random sample from a distribution :

$$\begin{aligned} f(x, \theta) &= \theta^x (1-\theta)^{1-x}; x = 0, 1, 0 < \theta < 1 \\ &= 0, \text{ elsewhere.} \end{aligned}$$

Show that  $Y_1 = x_1 + x_2 + \dots + x_n$ , is a sufficient statistic for  $\theta$ .

[Madras Univ. B.Sc., 1987]

(c) Let  $x_1, x_2, \dots, x_n$  denote a random sample from a population with p.d.f

$$f(x, \theta) = \theta x^{\theta-1}, \quad 0 < x < 1.$$

Show that  $Y = x_1 x_2 \dots x_n$ , is a sufficient statistic for  $\theta$ .

37. (a) Let  $X$  be a random sample of size one from a normal distribution  $N(0, \sigma^2)$ .

(i) Is  $X$  a sufficient statistic for  $\sigma^2$  ?

(ii) Is  $|X|$  a sufficient statistic for  $\sigma^2$  ?

(iii) Is  $X^2$  a sufficient statistic for  $\sigma^2$  ? (Gujarat Univ. B.Sc., 1992)

(b) Examine which of the following distributions admit sufficient estimators for their parameters :

(i)  $f(x, \theta) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1$

(ii)  $f(x, y, \rho) = \frac{1}{2\pi \sqrt{(1-\rho^2)}} \exp \left\{ -\frac{1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2) \right\}$

38. (a) Show that if a sufficient estimator exists, it is also the maximum likelihood estimator. Is the converse true ? Explain.

(b) Do the following distributions admit of sufficient estimators ?

(i)  $f(x, \theta) = \frac{1}{\theta}; k\theta \leq x \leq (k+1)\theta$ , where  $k$  is an integer.

(ii)  $f(x, \theta) = \frac{1+\theta}{(x+\theta)^2}, \quad 1 \leq x < \infty$

39. (a) Prove that if an unbiased estimator and a sufficient statistic exist for  $\psi(\theta)$  and the density function  $f(x, \theta)$  satisfies certain regularity conditions (to be stated by you), then the best unbiased estimate of  $\psi(\theta)$  is an explicit function of the sufficient statistic.

Examine if the following distribution admits a sufficient statistic for the parameter  $\theta$ .

$$f(x, \theta) = (1+\theta) x^\theta; \quad 0 \leq x \leq 1, \theta > 0$$

(b) Discuss if a sufficient statistic exists for the parameter  $\theta$ , in sampling from double exponential distribution with p.d.f.

$$f(x, \theta) = \frac{1}{2} \exp(-|x - \theta|), -\infty < x < \infty.$$

**Hint.** Proceed as in Example 15-17.

**Ans.** No sufficient estimator for  $\theta$  exists.

(c) Obtain jointly sufficient estimators for  $\alpha$  and  $\beta$  in a random sample  $X_1, X_2, \dots, X_n$  from the uniform population with p.d.f.

$$\begin{aligned} f(x, \alpha, \beta) &= \frac{1}{\beta - \alpha}, \quad \alpha \leq x \leq \beta \\ &= 0 \quad \text{otherwise} \end{aligned}$$

**Ans.**  $T_1 = X_{(1)}$  and  $T_2 = X_{(n)}$ , are jointly sufficient for  $\alpha$  and  $\beta$  respectively.

40. (a) Show that a necessary and sufficient condition for a statistic  $T$  to be sufficient for  $\theta$  is that the probability function  $f_\theta(x)$  should belong to an exponentially family.

(b) Let  $x_1, x_2, \dots, x_n$  be a random sample from a distribution with p.d.f.  $f(x : \theta) = e^{-(x-\theta)}, x \geq \theta, -\infty < \theta < \infty$ . Obtain a sufficient statistic for  $\theta$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1985]

41. Define a sufficient statistic. State and prove the Factorisation theorem on sufficiency. [Delhi Univ. B.Sc. (Stat. Hons.), 1986]

42. (a) Let  $(X_1, X_2, X_3)$  be a random sample from the probability mass function :  $P(X = x) = \theta^x (1 - \theta)^{1-x}, (x = 0, 1; 0 < \theta < 1)$ .

If  $t = X_1 + X_2 + X_3$ , show that the conditional distribution of the random sample given  $t = r$ , does not depend on  $\theta$ . Interpret this result in the light of sufficiency-concept.

(b) Let  $(X_1, X_2)$  be a random sample from a Poisson distribution with parameter  $\theta$ . Prove that  $t = X_1 + 2X_2$  is not sufficient for  $\theta$ .

(c) Let  $(X_1, X_2)$  be a random sample from  $N(\theta, 1)$ . If  $T = X_1 + X_2$  and  $U = X_2 - X_1$ , show that the conditional distribution of  $U$  given  $T = t$ , does not depend on  $\theta$ . Interpret this result in the light of sufficiency-concept.

(d) For a random sample  $X_i (i = 1, 2, \dots, n)$ , from an exponential distribution with p.d.f.

$$f(x, \theta) = \frac{1}{\theta} \exp\left[-\frac{x}{\theta}\right], x > 0, \theta > 0,$$

obtain an unbiased and sufficient estimator for  $\theta$ .

[Delhi Univ B.Sc. (Stat. Hons.) 1983, 1988]

43. Prove that under certain regularity conditions to be stated by you, the variance of an unbiased estimator  $T$  for  $\gamma(\theta)$ , satisfies the inequality

$$\text{Var}_\theta(T) \geq \frac{[\gamma'(\theta)]^2}{E_\theta\left[\frac{\partial \log f_\theta(X_1, X_2, \dots, X_n)}{\partial \theta}\right]^2}.$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986]

44. (a) If  $T$  is an unbiased estimator of a parameter  $\theta$ , based on a random sample of size  $n$ , prove that

$\text{Var}(T) \geq 1/[nI(\theta)]$ , where  $I(\theta)$  is the information function.

(b) Show that under certain regularity conditions, an unbiased estimate  $T$  of a parametric function  $\psi(\theta)$  attains a Cramer-Rao bound for the variance of unbiased estimator of  $\psi(\theta)$ , if and only if  $T$  satisfies the relation

$$\frac{\partial \log L}{\partial \theta} = \frac{n I(\theta)}{\psi'(\theta)} \{T - \psi(\theta)\}$$

where  $L$  is the likelihood function of a sample of  $n$  observations and

$$nI(\theta) = E\left(\frac{\partial \log L}{\partial \theta}\right)^2$$

What is the variance of  $T$  in such a case? Show that an estimator  $T$  satisfying the above relation is unique when it exists. Further a parametric function admitting such an estimator  $T$  is unique except for an additive and multiplicative constant. . (Meerut Univ. B.Sc., 1992)

45. (a) State and Prove Cramer-Rao Inequality.

(b) Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with p.d.f.

$$f(x, \theta) = \theta e^{-\theta x}; x > 0, \theta > 0.$$

Find Cramer-Rao lower bound for the variance of the unbiased estimator of  $\theta$ . [Delhi Univ. B.Sc. (Stat. Hons.), 1987]

46.  $f(x, \theta)$  is a probability density function and  $(x_1, x_2, \dots, x_n)$  is a random sample from it. Prove that if an unbiased minimum variance bound (MVB) estimator  $T$  exists, it must be of the form  $T = \theta + \lambda \sum_i \frac{\partial}{\partial \theta} \log f(x_i, \theta)$ , in which  $\lambda$  does not depend on sample values.

Show that the variance of  $T$  is  $\lambda$  and is given by

$$\frac{1}{\text{Var } T} = n E\left\{\frac{\partial^2}{\partial \theta^2} \log f(x_i, \theta)\right\}$$

Write a note on the connection between MVB estimators and sufficiency, giving example.

47. (a) Define Minimum Variance unbiased estimator and Minimum Variance Bound unbiased estimator and explain clearly the difference between them. Prove that minimum variance unbiased estimator is essentially unique.

(b) Verify that there exists an M.V.B. estimator for the parameter  $\theta$  of the distribution :  $f(x, \theta) = \frac{e^{-\theta} \theta^x}{x!}; x = 0, 1, 2, \dots$

and hence obtain the value of M.V.B. (Marathwada Univ. M.Sc., 1993)

(c) Show that there exists a parameter function  $\psi(\theta)$  in the case of the geometric distribution :

$$f(x, \theta) = (1 - \theta) \theta^x; x = 0, 1, 2, \dots; 0 < \theta < 1$$

such that there exists an M.V.B. unbiased estimator  $T$  of  $\psi(\theta)$ .

Obtain  $\psi(\theta), T$  and  $V(T)$ .

[Agra Univ. M.Sc., 1988]

48. (a) Define minimum variance unbiased estimator (MVUE). How is Cramer-Rao inequality useful in obtaining such an estimator? Derive this inequality.

(b) Obtain minimum variance unbiased estimator of  $\theta$  from a sample of  $n$  independent observations  $x_1, x_2, \dots, x_n$  drawn from the binomial  $B(N, \theta)$  population having probability function :

$$f(x; \theta) = {}^N C_x \theta^x (1 - \theta)^{N-x}, x = 0, 1, 2, \dots, N.$$

Also obtain variance of this estimator of  $\theta$ .

49. (a) If  $b(\theta)$  is the bias in the estimator  $T$  of  $\theta$ , then show that (under conditions to be stated by you),

$$E(T - \theta)^2 \geq \frac{(1 + b'(\theta))^2}{I(\theta)} + (b(\theta))^2,$$

where  $I(\theta)$  is the information on  $\theta$  supplied by a sample of  $n$  observations.

(b) Prove the following result :

$$\int_{-\infty}^{\infty} (x - \theta)^2 \cdot g(x, \theta) dx \int_{-\infty}^{\infty} \left( \frac{\partial \log g}{\partial \theta} \right)^2 g(x, \theta) dx \geq \left( \frac{d\psi}{d\theta} \right)^2$$

where  $g(x, \theta)$  is the frequency function in  $x$  having the first moment  $\psi(\theta)$  and finite second moment. Discuss when the equality sign holds.

50. For the gamma distribution

$$f(x, \theta) = \frac{1}{\theta^p \Gamma p} x^{p-1} \exp(-x/\theta); 0 \leq x < \infty, \theta > 0, p \text{ (known)},$$

find the expectation of  $X^2$ . Use it to obtain an unbiased estimator  $T$  of  $\theta^2$ . Find  $V(T)$ .

Evaluate Fisher's information function  $I(\theta)$  about  $\theta^2$  and verify the truth of the inequality :

$$V(T) > [n I(\theta^2)]^{-1}$$

51. State and prove Rao-Blackwell theorem and explain its significance in the theory of point estimation.

Let  $x_1, x_2, \dots, x_n$  be a random sample from Poisson distribution with parameter  $\lambda$ . Obtain Cramer-Rao lower bound to the variance of an unbiased estimator for  $\lambda$ . Hence find the M.V.U.E. for  $\lambda$ .

[Delhi Univ. M.Sc., (Maths.), 1990]

52. State and prove Rao-Blackwell theorem and explain its significance in point estimation.

Let  $X_1, X_2, \dots, X_n$  be a random sample from a rectangular distribution with p.d.f.

$$f(x, \theta) = 1/\theta, 0 \leq x \leq \theta.$$

Find MVU estimators of  $\theta$  and  $3\theta + 5$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1987]

53. Define completeness of a statistic  $T$ . Let  $X_1, X_2, \dots, X_n$  be a random sample from uniform population  $U[0, \theta]$ . Obtain sufficient statistic for  $\theta$ . Show that it is complete. Hence obtain MVU estimator for  $\theta$ .

[Delhi Univ B.Sc. (Stat. Hons.), 1988]

54. Define a complete sufficient statistic.

If  $T$  is a complete sufficient statistic for  $\gamma(\theta)$ , and  $E[\phi(T)] = \gamma(\theta)$ , then show that  $\phi(T)$  is the unique MVUE of  $\gamma(\theta)$ .

Use this property and obtain MVU estimator of  $\theta$  based on a random sample  $X_1, X_2, \dots, X_n$  from the distribution with p.m.f.

$$f(x, \theta) = \begin{cases} \theta^x (1-\theta)^{1-x}, & x=0, 1 \\ 0, & \text{elsewhere} \end{cases}$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

55. Show that the family  $\{f(x, \theta), \theta \in (0, 1)\}$  with

$$f(x, \theta) = {}^2C_x \theta^x (1-\theta)^{2-x}, \quad x=0, 1, 2,$$

is complete. [Delhi Univ. B.Sc. (Stat. Hons.), 1993]

56. Let  $X_1, X_2, \dots, X_n$  be a random sample from

$$f_\theta(x) = \frac{1}{\theta}, \quad 0 < x < \theta \text{ for all } \theta \in \Theta$$

Show that  $X_{(n)} = \max(X_1, X_2, \dots, X_n)$  is sufficient for  $\theta$  and  $\frac{(n+1)}{n} X_{(n)}$  is an unbiased estimator for  $\theta$ .

Comment on the result.

[Agra Univ. M.Sc., 1988]

57. Let the random variables  $X$  and  $Y$  have the joint p.d.f.

$$f(x, y) = \frac{2}{\theta^2} \exp\left[-\frac{(x+y)}{\theta}\right], \quad 0 < x < y < \infty$$

and zero elsewhere.

(a) Show that :  $E(Y | x) = x + \theta$

Obtain the expected value of  $X + \theta$  and compare the variance of  $X + \theta$  with that of  $Y$ . [Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986]

(b) Show that :  $E(Y) = \frac{3}{2}\theta$ ,  $\text{Var}(Y) = \frac{5}{4}\theta^2$ .

[Madras Univ. B.Sc., 1988]

58. (a) A random sample of size  $n$  is drawn from a Poisson population with parameters  $\lambda$ . Obtain the minimum variance unbiased estimator of  $\lambda$ .

[Delhi Univ. M.A. (Eco.), 1992]

(b) Establish a necessary and sufficient condition for an unbiased estimator to be an MVU estimator.

Let  $X_1, X_2, \dots, X_n$  be a random sample from a Poisson distribution with parameter  $\theta$ . Find an MVU estimator for  $\gamma(\theta) = e^{-\theta} \theta^4 / 24$ .

59. Define sufficiency of an estimator

Let  $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$  be the order statistics of a random sample of size 5 from the uniform distribution with p.d.f.

$$f(x, \theta) = \begin{cases} \frac{1}{\theta}; & 0 < x < \theta, 0 < \theta < \infty \\ 0, & \text{elsewhere} \end{cases}$$

Show that  $2Y_3$  is an unbiased estimator of  $\theta$ . Find the conditional expectation  $E[2Y_3 | Y_5] = \phi(Y_5)$ , say. Compare the variances of  $2Y_3$  and  $\phi(Y_5)$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

60. Let  $X_1, X_2, \dots, X_n$  be a random sample from the Bernoulli population with parameter  $\theta$ ,  $0 < \theta < 1$ . Obtain a sufficient statistic for  $\theta$  and show that it is complete. Hence obtain MVU estimator of  $\theta$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

61. Show that  $T = \sum_{i=1}^n X_i$ , is a complete sufficient statistic for the parameter  $\theta$  in a random sample  $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n$  drawn from the population with p.d.f.

$$(a) f(x, \theta) = \theta^x (1-\theta)^{1-x}; x = 0, 1 \\ = 0, \text{ elsewhere}$$

$$(b) f(x, \theta) = \begin{cases} e^{-\theta} \theta^x / x!, & x = 0, 1, 2, \dots \\ 0, & \text{elsewhere} \end{cases}$$

62. If  $X_1, X_2, \dots, X_n$  is a random sample from  $N(\mu, \sigma^2)$ , show that :

(a)  $T = \bar{X}$ , is complete sufficient statistic for  $\mu$ , ( $-\infty < \mu < \infty$ ), when  $\sigma^2$  is known.

(b)  $T = \sum_{i=1}^n (X_i - \mu)^2$ , is complete sufficient statistic for  $\sigma^2$ , ( $0 < \sigma^2 < \infty$ ), when  $\mu$  is known.

**15.10. Methods of Estimation.** So far we have been discussing the requisites of a good estimator. Now we shall briefly outline some of the important methods for obtaining such estimators. Commonly used methods are

- (i) Method of Maximum Likelihood Estimation.
- (ii) Method of Minimum Variance.
- (iii) Method of Moments.
- (iv) Method of Least Squares.
- (v) Method of Minimum Chi-square
- (vi) Method of Inverse Probability.

In the following sections, we shall discuss briefly the first four methods only.

**15.11. Method of Maximum Likelihood Estimation.** From theoretical point of view, the most general method of estimation known is the method of Maximum Likelihood Estimators (M.L.E.) which was initially formulated by C.F. Gauss but as a general method of estimation was first introduced by Prof. R.A. Fisher and later on developed by him in a series of papers. Before introducing the method we will first define *Likelihood Function*.

**Likelihood Function.** *Definition.* Let  $x_1, x_2, \dots, x_n$  be a random sample of size  $n$  from a population with density function  $f(x, \theta)$ . Then the likelihood function of the sample values  $x_1, x_2, \dots, x_n$ , usually denoted by  $L = L(\theta)$  is their joint density function, given by

$$L = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta) = \prod_{i=1}^n f(x_i, \theta). \quad \dots(15.53)$$

$L$  gives the relative likelihood that the random variables assume a particular set of values  $x_1, x_2, \dots, x_n$ . For a given sample  $x_1, x_2, \dots, x_n$ ,  $L$  becomes a function of the variable  $\theta$ , the parameter.

The principle of maximum likelihood consists in finding an estimator for the unknown parameter  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ , say, which maximises the likelihood function  $L(\theta)$  for variations in parameter i.e., we wish to find  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$  so that

$$L(\hat{\theta}) > L(\theta) \quad \forall \theta \in \Theta$$

$$\text{i.e., } L(\hat{\theta}) = \text{Sup } L(\theta) \quad \forall \theta \in \Theta.$$

Thus if there exists a function  $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$  of the sample values which maximises  $L$  for variations in  $\theta$ , then  $\hat{\theta}$  is to be taken as an estimator of  $\theta$ .  $\hat{\theta}$  is usually called *Maximum Likelihood Estimator (M.L.E.)*. Thus  $\hat{\theta}$  is the solution, if any, of

$$\frac{\partial L}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial^2 L}{\partial \theta^2} < 0 \quad \dots(15-54)$$

Since  $L > 0$ , and  $\log L$  is a non-decreasing function of  $L$ ;  $L$  and  $\log L$  attain their extreme values (maxima or minima) at the same value of  $\hat{\theta}$ . The first of the two equations in (15-54) can be rewritten as

$$\frac{1}{L} \cdot \frac{\partial L}{\partial \theta} = 0 \Rightarrow \frac{\partial \log L}{\partial \theta} = 0, \quad \dots(15-54a)$$

a form which is much more convenient from practical point of view.

If  $\theta$  is vector valued parameter, then  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ , is given by the solution of simultaneous equations :

$$\frac{\partial}{\partial \theta_i} \log L = \frac{\partial}{\partial \theta_i} \log L(\theta_1, \theta_2, \dots, \theta_k) = 0; \quad i = 1, 2, \dots, k \quad \dots(15-54b)$$

Equations (15-54a) and (15-54b) are usually referred to as the *Likelihood Equations* for estimating the parameters.

**Remark.** For the solution  $\hat{\theta}$  of the likelihood equations, we have to see that the second derivative of  $L$  w.r. to  $\theta$  is negative. If  $\theta$  is vector valued, then for  $L$  to be maximum, the matrix of derivatives

$$\left( \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right)_{\theta=\hat{\theta}} \text{ should be negative definite.}$$

### 15.11.1. Properties of Maximum Likelihood Estimators.

We make the following assumptions, known as the *Regularity Conditions*:

(i) The first and second order derivatives, viz.,  $\frac{\partial \log L}{\partial \theta}$  and  $\frac{\partial^2 \log L}{\partial \theta^2}$  exist and are continuous functions of  $\theta$  in a range  $R$  (including the true value  $\theta_0$  of the parameter) for almost all  $x$ . For every  $\theta$  in  $R$

$$\left| \frac{\partial}{\partial \theta} \log L \right| < F_1(x) \quad \text{and} \quad \left| \frac{\partial^2}{\partial \theta^2} \log L \right| < F_2(x)$$

where  $F_1(x)$  and  $F_2(x)$  are integrable functions over  $(-\infty, \infty)$ .

(ii) The third order derivative  $\frac{\partial^2}{\partial \theta^3} \log L$  exists such that

$$\left| \frac{\partial^3}{\partial \theta^3} \cdot \log L \right| < M(x)$$

where  $E[M(x)] < K$ , a positive quantity.

(iii) For every  $\theta$  in  $R$ ,

$$E \left( -\frac{\partial^2}{\partial \theta^2} \log L \right) = \int_{-\infty}^{\infty} \left( -\frac{\partial^2}{\partial \theta^2} \log L \right) L dx \\ = I(\theta),$$

is finite and non-zero.

(iv) The range of integration is independent of  $\theta$ . But if the range of integration depends on  $\theta$ , then  $f(x, \theta)$  vanishes at the extremes depending on  $\theta$ .

This assumption is to make the differentiation under the integral sign valid.

Under the above assumptions M.L.E. possesses a number of important properties, which will be stated in the form of theorems.

**Theorem 15-11. (Cramer-Rao Theorem).** "With probability approaching unity as  $n \rightarrow \infty$ , the likelihood equation  $\frac{\partial}{\partial \theta} \log L = 0$ , has a solution which converges in probability to the true value  $\theta_0$ ". In other words M.L.E.'s are consistent.

**Remark.** MLE's are always consistent estimators but need not be unbiased. For example in sampling from  $N(\mu, \sigma^2)$  population, [c.f. Example 15-31],

$\text{MLE}(\mu) = \bar{x}$  (sample mean), which is both unbiased and consistent estimator of  $\mu$ .

$\text{MLE}(\sigma^2) = s^2$  (sample variance), which is consistent but not unbiased estimator of  $\sigma^2$ .

**Theorem 15-12. (Hazard Bazar's Theorem).** Any consistent solution of the likelihood equation provides a maximum of the likelihood with probability tending to unity as the sample size ( $n$ ) tends to infinity.

**Theorem 15-13. (Asymptotic Normality of MLE's).** A consistent solution of the likelihood equation is asymptotically normally distributed about the true value  $\theta_0$ . Thus,  $\hat{\theta}$  is asymptotically  $N\left(\theta_0, \frac{I}{I(\theta_0)}\right)$  as  $n \rightarrow \infty$ .

**Remark.** Variance of M.L.E. is given by

$$V(\hat{\theta}) = \frac{1}{I(\theta)} = \frac{1}{\left[ E \left( -\frac{\partial^2}{\partial \theta^2} \log L \right) \right]} \quad \dots(15-55)$$

**Theorem 15-14.** If M.L.E. exists, it is the most efficient in the class of such estimators.

**Theorem 15-15.** If a sufficient estimator exists, it is a function of the Maximum Likelihood Estimator.

**Proof.** If  $t = t(x_1, x_2, \dots, x_n)$  is a sufficient estimator of  $\theta$ , then Likelihood Function can be written as (c.f. Theorem 15-7)

$$L = g(t, \theta) h(x_1, x_2, x_3, \dots, x_n | t)$$

where  $g(t, \theta)$  is the density function of  $t$  and  $h(x_1, x_2, \dots, x_n | t)$  is the density function of the sample, given  $t$ , and is independent of  $\theta$ .

$$\therefore \log L = \log g(t, \theta) + \log h(x_1, x_2, \dots, x_n | t)$$

Differentiating w.r.t.  $\theta$ , we get

$$\frac{\partial \log L}{\partial \theta} = \frac{\partial}{\partial \theta} \log g(t, \theta) = \psi(t, \theta), \text{ (say)}, \quad \dots(15-56)$$

which is a function of  $t$  and  $\theta$  only.

M.L.E. is given by

$$\frac{\partial \log L}{\partial \theta} = 0 \Rightarrow \psi(t, \theta) = 0$$

$$\therefore \hat{\theta} = \eta(t) = \text{Some function of sufficient statistic.}$$

$$\Rightarrow \hat{t} = \psi(\theta) = \text{Some function of M.L.E.}$$

Hence the theorem.

**Remark.** This theorem is quite helpful in finding if a sufficient estimator exists or not.

If  $\frac{\partial}{\partial \theta} \log L$  can be expressed in the form (15-56), i.e., as a function of a statistic and parameter alone, then the statistic is regarded as a sufficient estimator of the parameter. If  $\frac{\partial}{\partial \theta} \log L$  cannot be expressed in the form (15-56), no sufficient estimator exists in that case.

**Theorem 15-16.** If for a given population with p.d.f.  $f(x, \theta)$ , an MVB estimator  $T$  exists for  $\theta$ , then the likelihood equation will have a solution equal to the estimator  $T$ .

**Proof.** Since  $T$  is an MVB estimator of  $\theta$ , we have [c.f. (15-40)],

$$\frac{\partial}{\partial \theta} \log L = \frac{T - \theta}{\lambda(\theta)} = (T - \theta) A(\theta)$$

MLE for  $\theta$  is the solution of the likelihood equation

$$\frac{\partial}{\partial \theta} \log L = 0 \Rightarrow \hat{\theta} = T$$

as required.

**Theorem 15-17. (Invariance Property of MLE).** If  $T$  is the MLE of  $\theta$  and  $\psi(\theta)$  is one to one function of  $\theta$ , then  $\psi(T)$  is the MLE of  $\psi(\theta)$ .

**Example 15-31.** In random sampling from normal population  $N(\mu; \sigma^2)$ , find the maximum likelihood estimators for

- (i)  $\mu$  when  $\sigma^2$  is known,
- (ii)  $\sigma^2$  when  $\mu$  is known, and

(iii) the simultaneous estimation of  $\mu$  and  $\sigma^2$ .

[Madras Univ. B.Sc. Sept., 1987]

**Solution.**  $X \sim N(\mu, \sigma^2)$  then

$$\begin{aligned} L &= \prod_{i=1}^n \left[ \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right\} \right] \\ &= \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^n \exp \left\{ -\sum_{i=1}^n (x_i - \mu)^2 / 2\sigma^2 \right\} \end{aligned}$$

$$\log L = -\frac{n}{2} \log (2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

**Case (i).** When  $\sigma^2$  is known, the likelihood equation for estimating  $\mu$  is

$$\frac{\partial}{\partial \mu} \log L = 0 \Rightarrow -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(x_i - \mu)(-1) = 0$$

or  $\sum_{i=1}^n (x_i - \mu) = 0 \Rightarrow \sum_{i=1}^n x_i - n\mu = 0$

$$\Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \quad \dots (*)$$

Hence M.L.E. for  $\mu$  is the sample mean  $\bar{x}$ .

**Case (ii).** When  $\mu$  is known, the likelihood equation for estimating  $\sigma^2$  is

$$\frac{\partial}{\partial \sigma^2} \log L = 0 \Rightarrow -\frac{n}{2} \times \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

$$\Rightarrow n - \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = 0, \quad i.e., \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad \dots (**)$$

**Case (iii).** The likelihood equations for simultaneous estimation of  $\mu$  and  $\sigma^2$  are

$$\frac{\partial}{\partial \mu} \log L = 0 \text{ and } \frac{\partial}{\partial \sigma^2} \log L = 0, \text{ thus giving}$$

$$\hat{\mu} = \bar{x} \quad \text{[From (*)]}$$

and  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad \text{[From (**)]}$

$$= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2, \text{ the sample variance.}$$

**Important Note.** It may be pointed out here that though

$$\left. \begin{aligned} E(\hat{\mu}) &= E(\bar{x}) = \mu \\ E(\hat{\sigma}^2) &= E(s^2) \neq \sigma^2 \end{aligned} \right\} \quad (c.f. \S \ 12-12)$$

Hence the maximum likelihood estimators (M.L.E.s.) need not necessarily be unbiased.

**Remark.** Since M.L.E. is the most efficient, we conclude that in sampling from a normal population, the sample mean  $\bar{x}$  is the most efficient estimator of the population mean  $\mu$ .

**Example 15-32.** Prove that the maximum likelihood estimate of the parameter  $\alpha$  of a population having density function :

$$\frac{2}{\alpha^2}(\alpha - x), \quad 0 < x < \alpha$$

for a sample of unit size is  $2x$ ,  $x$  being the sample value. Show also that the estimate is biased. [Burdwan Univ. B.Sc. (Maths. Hons.), 1991]

**Solution.** For a random sample of unit size ( $n = 1$ ), the likelihood function is :

$$L(\alpha) = f(x, \alpha) = \frac{2}{\alpha^2}(\alpha - x); \quad 0 < x < \alpha$$

Likelihood equation gives :

$$\begin{aligned} \frac{d}{d\alpha} \log L &= \frac{d}{d\alpha} [\log 2 - 2 \log \alpha + \log (\alpha - x)] = 0 \\ \Rightarrow -\frac{2}{\alpha} + \frac{1}{\alpha - x} &= 0 \quad \Rightarrow \quad 2(\alpha - x) - \alpha = 0 \quad \Rightarrow \quad \alpha = 2x \end{aligned}$$

Hence MLE of  $\alpha$  is given by  $\hat{\alpha} = 2x$ .

$$\begin{aligned} E(\hat{\alpha}) &= E(2X) = 2 \int_0^\alpha x \cdot f(x, \alpha) dx \\ &= \frac{4}{\alpha^2} \int_0^\alpha x(\alpha - x) dx = \frac{4}{\alpha^2} \left| \frac{\alpha x^2}{2} - \frac{x^3}{3} \right|_0^\alpha = \frac{2}{3} \alpha \end{aligned}$$

Since  $E(\hat{\alpha}) \neq \alpha$ ,  $\hat{\alpha} = 2x$  is not an unbiased estimate of  $\alpha$ .

**Example 15-33.** (a) Find the maximum likelihood estimate for the parameter  $\lambda$  of a Poisson distribution on the basis of a sample of size  $n$ . Also find its variance.

(b) Show that the sample mean  $\bar{x}$ , is sufficient for estimating the parameter  $\lambda$  of the Poisson distribution.

**Solution.** The probability function of the Poisson distribution with parameter  $\lambda$  is given by

$$P(X=x) = f(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, 2, \dots$$

Likelihood function of random sample  $x_1, x_2, \dots, x_n$  of  $n$  observations from this population is

$$L = \prod_{i=1}^n f(x_i, \lambda) = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}}{x_1! x_2! \dots x_n!}$$

$$\begin{aligned}\therefore \log L &= -n\lambda + \left( \sum_{i=1}^n x_i \right) \log \lambda - \sum_{i=1}^n \log(x_i !) \\ &= -n\lambda + n\bar{x} \log \lambda - \sum_{i=1}^n \log(x_i !)\end{aligned}$$

The likelihood equation for estimating  $\lambda$  is

$$\frac{\partial}{\partial \lambda} \log L = 0 \Rightarrow -n + \frac{n\bar{x}}{\lambda} = 0 \Rightarrow \lambda = \bar{x}$$

Thus the M.L.E. for  $\lambda$  is the sample mean  $\bar{x}$ .

The variance of the estimate is given by

$$\begin{aligned}V(\hat{\lambda}) &= E\left[-\frac{\partial^2}{\partial \lambda^2}(\log L)\right] \quad [c.f. (15-55)] \\ &= E\left[-\frac{\partial}{\partial \lambda}\left(-n + \frac{n\bar{x}}{\lambda}\right)\right] = E\left[-\left(-\frac{n\bar{x}}{\lambda^2}\right)\right] = \frac{n}{\lambda^2} E(\bar{x}) = \frac{n}{\lambda} \\ \therefore V(\hat{\lambda}) &= \lambda/n\end{aligned}$$

(b) For the Poisson distribution with parameter  $\lambda$ , we have

$$\begin{aligned}\frac{\partial}{\partial \lambda} \log L &= -n + \frac{n\bar{x}}{\lambda} \\ &= n\left(\frac{\bar{x}}{\lambda} - 1\right) = \psi(\bar{x}, \lambda), \text{ a function of } \bar{x} \text{ and } \lambda \text{ only.}\end{aligned}$$

Hence (c.f. Remark Theorem 15-15),  $\bar{x}$  is sufficient for estimating  $\lambda$ .

**Example 15-34.** Let  $x_1, x_2, \dots, x_n$  denote random sample of size  $n$  from a uniform population with p.d.f.

$$f(x, \theta) = 1 ; \theta - \frac{1}{2} \leq x \leq \theta + \frac{1}{2}, -\infty < \theta < \infty$$

Obtain M.L.E. for  $\theta$ .

[Delhi Univ. M.C.A., 1987]

**Solution.** Here

$$\begin{aligned}L = L(\theta; x_1, x_2, \dots, x_n) &= 1, \theta - \frac{1}{2} \leq x_i \leq \theta + \frac{1}{2} \\ &= 0, \text{ elsewhere}\end{aligned}$$

If  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  is the ordered sample then

$$\theta - \frac{1}{2} \leq x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} \leq \theta + \frac{1}{2}$$

Thus  $L$  attains the maximum if

$$\begin{aligned}\theta - \frac{1}{2} \leq x_{(1)} \quad \Lambda \quad x_{(n)} \leq \theta + \frac{1}{2} \\ \Rightarrow \quad \theta \leq x_{(1)} + \frac{1}{2} \quad \Lambda \quad x_{(n)} - \frac{1}{2} \leq \theta\end{aligned}$$

Hence every statistic  $t = t(x_1, x_2, \dots, x_n)$  such that

$$x_{(n)} - \frac{1}{2} \leq t(x_1, x_2, \dots, x_n) \leq x_{(1)} + \frac{1}{2}$$

provides an M.L.E. for  $\theta$ .

**Remark.** This example illustrates that M.L.E. for a parameter need not be unique.

**Example 15-35.** Find the M.L.E. of the parameters  $\alpha$  and  $\lambda$ , ( $\lambda$  being large), of the distribution :

$$f(x; \alpha, \lambda) = \frac{1}{\Gamma(\lambda)} \left( \frac{\lambda}{\alpha} \right)^{\lambda} e^{-\lambda x / \alpha} x^{\lambda-1}; 0 \leq x < \infty, \lambda > 0$$

You may use that for large values of  $\lambda$ ,

$$\psi(\lambda) = \frac{\partial}{\partial \lambda} \log \Gamma(\lambda) = \log \lambda - \frac{1}{2\lambda}$$

$$\text{and } \psi'(\lambda) = \frac{1}{\lambda} + \frac{1}{2\lambda^2}$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1985]

**Solution.** Let  $x_1, x_2, \dots, x_n$  be a random sample of size  $n$  from the given population. Then

$$L = \prod_{i=1}^n f(x_i; \alpha, \lambda) = \left( \frac{1}{\Gamma(\lambda)} \right)^n \cdot \left( \frac{\lambda}{\alpha} \right)^{n\lambda} \cdot \exp \left[ -\frac{\lambda}{\alpha} \sum_{i=1}^n x_i \right] \cdot \prod_{i=1}^n (x_i^{\lambda-1})$$

$$\therefore \log L = -n \log \Gamma(\lambda) + n\lambda(\log \lambda - \log \alpha) - \frac{\lambda}{\alpha} \sum_{i=1}^n x_i + (\lambda - 1) \sum_{i=1}^n \log x_i$$

If  $G$  is the geometric mean of  $x_1, x_2, \dots, x_n$ , then

$$\log G = \frac{1}{n} \sum_{i=1}^n \log x_i \Rightarrow n \log G = \sum_{i=1}^n \log x_i$$

$$\therefore \log L = -n \log \Gamma(\lambda) + n\lambda(\log \lambda - \log \alpha) - \frac{\lambda}{\alpha} n\bar{x} + (\lambda - 1) n \log G$$

where  $G$  is independent of  $\lambda$  and  $\alpha$ .

The likelihood equations for the simultaneous estimation of  $\alpha$  and  $\lambda$  are :

$$\frac{\partial}{\partial \alpha} \log L = 0 \dots (1) \quad \text{and} \quad \frac{\partial}{\partial \lambda} \log L = 0 \dots (2)$$

(1) gives

$$-\frac{n\lambda}{\alpha} + \frac{\lambda}{\alpha^2} \cdot n\bar{x} = 0 \Rightarrow -1 + \frac{\bar{x}}{\alpha} = 0 \Rightarrow \hat{\alpha} = \bar{x} \dots (*)$$

(2) gives (for large values of  $\lambda$ ),

$$-n \left( \log \lambda - \frac{1}{2\lambda} \right) + n \left[ 1 \cdot (\log \lambda - \log \alpha) + \lambda \cdot \frac{1}{\lambda} \right] - \frac{n\bar{x}}{\alpha} + n \log G = 0$$

$$\Rightarrow \frac{1}{2\lambda} + \left( 1 - \log \alpha + \log G - \frac{\bar{x}}{\alpha} \right) = 0$$

$$\Rightarrow 1 + 2\lambda (\log G - \log \bar{x}) = 0$$

[From (\*)]

$$\Rightarrow 1 - 2\lambda \log \left( \frac{\bar{x}}{G} \right) = 0, \text{ i.e., } \hat{\lambda} = \frac{1}{2 \log (\bar{x}/G)}$$

Hence the M.L.E.s for  $\alpha$  and  $\lambda$  are given by

$$\hat{\alpha} = \bar{x} \quad \text{and} \quad \hat{\lambda} = \frac{1}{2 \log (\bar{x}/G)},$$

**Example 15-36.** In sampling from a power series distribution with p.d.f.  $f(x, \theta) = a_x \theta^x / \psi(\theta); x = 0, 1, 2, \dots$

where  $a_x$  may be zero for some  $x$ , show that MLE of  $\theta$  is a root of the equation

$$\bar{X} = \frac{\theta \psi'(\theta)}{\psi(\theta)} = \mu(\theta), \quad \dots (*)$$

where  $\mu(\theta) = E(X)$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

**Solution.** Likelihood function is given by :

$$L = \prod_{i=1}^n f(x_i, \theta) = \prod_{i=1}^n \left[ \frac{a_{x_i} \theta^{x_i}}{\psi(\theta)} \right] = \left[ \prod_{i=1}^n a_{x_i} \right] \frac{\theta^{\sum x_i}}{[\psi(\theta)]^n}$$

$$\Rightarrow \log L = \sum_{i=1}^n \log a_{x_i} + \log \theta \cdot \sum_{i=1}^n x_i - n \log \psi(\theta)$$

Likelihood equation for estimating  $\theta$  gives :

$$\frac{\partial}{\partial \theta} \log L = 0 = \frac{\sum x_i}{\theta} - \frac{n \psi'(\theta)}{\psi(\theta)}$$

$$\Rightarrow \bar{X} = \frac{\sum x_i}{n} = \frac{\theta \psi'(\theta)}{\psi(\theta)} = \mu(\theta), \text{ (say).} \quad \dots (**)$$

Hence MLE of  $\theta$  is a root of equation (\*).

We have :

$$E(X) = \sum_{x=0}^{\infty} x f(x, \theta) = \sum_{x=0}^{\infty} \left[ x \left\{ \frac{a_x \theta^x}{\psi(\theta)} \right\} \right] \quad \dots (**)$$

$$\sum_{x=0}^{\infty} f(x, \theta) = 1 \Rightarrow \sum_{x=0}^{\infty} \frac{a_x \theta^x}{\psi(\theta)} = 1 \Rightarrow \sum_{x=0}^{\infty} a_x \theta^x = \psi(\theta)$$

Differentiating w.r. to  $\theta$ , we get

$$\sum_x [a_x \cdot x \theta^{x-1}] = \psi'(\theta)$$

$$\Rightarrow \sum_x \left[ a_x \cdot \frac{x \theta^x}{\psi(\theta)} \right] = \frac{\theta \psi'(\theta)}{\psi(\theta)}$$

$$\Rightarrow E(X) = \mu(\theta) = \bar{X},$$

[From (\*\*\*) and (\*)]

as required.

**Example 15-37.** (a) Let  $x_1, x_2, \dots, x_n$  be a random sample from the uniform distribution with p.d.f.

$$f(x, \theta) = \begin{cases} \frac{1}{\theta}, & 0 < x < \infty, \theta > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Obtain the maximum likelihood estimator for  $\theta$ .

[*Lucknow Univ. B.Sc., 1992*]

(b) Obtain the M.L.E.s. for  $\alpha$  and  $\beta$  for the rectangular population

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha}, & \alpha < x < \beta \\ 0, & \text{elsewhere} \end{cases}$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989; Gujarat Univ. B.Sc. 1992*]

**Solution.** (a) Here

$$L = \prod_{i=1}^n f(x_i, \theta) = \frac{1}{\theta} \cdot \frac{1}{\theta} \cdots \frac{1}{\theta} = \left( \frac{1}{\theta} \right)^n \quad \dots (*)$$

Likelihood equation, viz.,  $\frac{\partial}{\partial \theta} \log L = 0$ , gives

$$\frac{\partial}{\partial \theta} (-n \log \theta) = 0 \Rightarrow -\frac{n}{\theta} = 0 \Rightarrow \hat{\theta} = \infty,$$

obviously an absurd result.

In this case we locate M.L.E. as follows :

We have to choose  $\theta$  so that  $L$  in (\*) is maximum. Now  $L$  is maximum if  $\theta$  is minimum.

Let  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  be the *ordered* sample of  $n$  independent observations from the given population so that

$$0 \leq x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} \leq \theta \Rightarrow \theta \geq x_{(n)}$$

Since the minimum value of  $\theta$  consistent with the sample is  $x_{(n)}$ , the largest sample observation,  $\hat{\theta} = x_{(n)}$ .

$\therefore$  M.L.E. for  $\theta = x_{(n)} =$  The largest sample observation.

(b) Here

$$L = \left( \frac{1}{\beta - \alpha} \right)^n \quad \dots (**)$$

$$\therefore \log L = -n \log (\beta - \alpha)$$

The likelihood equations for  $\alpha$  and  $\beta$  give

$$\left. \begin{aligned} \frac{\partial}{\partial \alpha} \log L = 0 &= \frac{n}{\beta - \alpha} \\ \frac{\partial}{\partial \beta} \log L = 0 &= \frac{-n}{\beta - \alpha} \end{aligned} \right\}$$

and

Each of these equations gives  $\beta - \alpha = \infty$ , an obviously negative result. So, we find M.L.E.s for  $\alpha$  and  $\beta$  by some other means.

Now  $L$  in  $(**)$  is maximum if  $(\beta - \alpha)$  is minimum, i.e., if  $\beta$  takes the minimum possible value and  $\alpha$  takes the maximum possible value.

As in part (a), if  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  is an ordered random sample from this population, then  $\alpha \leq x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} \leq \beta$ . Thus  $\beta \geq x_{(n)}$  and  $\alpha \leq x_{(1)}$ . Hence the minimum possible value of  $\beta$  consistent with the sample is  $x_{(n)}$  and the maximum possible value of  $\alpha$  consistent with the sample is  $x_{(1)}$ . Hence  $L$  is maximum if  $\beta = x_{(n)}$  and  $\alpha = x_{(1)}$ .

$\therefore$  M.L.E. for  $\alpha$  and  $\beta$  are given by

$$\hat{\alpha} = x_{(1)} = \text{The smallest sample observation}$$

and  $\hat{\beta} = x_{(n)} = \text{The largest sample observation.}$

**Example 15.38.** State as precisely as possible the properties of the M.L.E. Obtain the M.L.E.s. of  $\alpha$  and  $\beta$  for a random sample from the exponential population

$$f(x; \alpha, \beta) = y_0 e^{-\beta(x-\alpha)}, \alpha \leq x \leq \infty, \beta > 0$$

$y_0$  being a constant.

**Solution.** Here first of all we shall determine the constant  $y_0$  from the consideration that the total area under a probability curve is unity.

$$\therefore y_0 \int_{\alpha}^{\infty} \exp[-\beta(x-\alpha)] dx = 1$$

$$\text{or } y_0 \left| \frac{e^{-\beta(x-\alpha)}}{-\beta} \right|_{\alpha}^{\infty} = 1 \Rightarrow -\frac{y_0}{\beta}(0-1) = 1 \Rightarrow y_0 = \beta$$

$$\therefore f(x; \alpha, \beta) = \beta e^{-\beta(x-\alpha)}, \alpha \leq x < \infty$$

If  $x_1, x_2, \dots, x_n$  is a random sample of  $n$  observations from this population, then

$$L = \prod_{i=1}^n f(x_i; \alpha, \beta) = \beta^n \exp \left\{ -\beta \sum_{i=1}^n (x_i - \alpha) \right\} = \beta^n e^{-n\beta(\bar{x} - \alpha)}$$

$$\therefore \log L = n \log \beta - n\beta(\bar{x} - \alpha) \quad ... (*)$$

The likelihood equations for estimating  $\alpha$  and  $\beta$  give

$$\frac{\partial}{\partial \alpha} \log L = 0 = n\beta \quad ... (**)$$

$$\text{and } \frac{\partial}{\partial \beta} \log L = 0 = \frac{n}{\beta} - n(\bar{x} - \alpha) \quad ... (***)$$

Equation  $(**)$  gives  $\beta = 0$ , which is obviously inadmissible and this on substitution in  $(***)$  gives  $\alpha = \infty$ , a nugatory result. Thus the likelihood equations fail to give us valid estimates of  $\alpha$  and  $\beta$  and we try to locate M.L.E.s. for  $\alpha$  and  $\beta$  by maximising  $L$  directly.

$L$  is maximum  $\Rightarrow \log L$  is maximum.

From  $(*)$ ,  $\log L$  is maximum (for any value of  $\beta$ ), if  $(\bar{x} - \alpha)$  is minimum, which is so if  $\alpha$  is maximum.

If  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  is ordered sample from this population then

$$\alpha \leq x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} < \infty,$$

so that the maximum value of  $\alpha$  consistent with the sample is  $x_{(1)}$ , the smallest sample observation, i.e.,

$$\hat{\alpha} = x_{(1)}$$

Consequently, (\*\*\*\*) gives

$$\frac{1}{\hat{\beta}} = \bar{x} - \hat{\alpha} = \bar{x} - x_{(1)} \Rightarrow \hat{\beta} = \frac{1}{\bar{x} - x_{(1)}}$$

Hence M.L.Es. for  $\alpha$  and  $\beta$  are given by

$$\hat{\alpha} = x_{(1)} \text{ and } \hat{\beta} = \frac{1}{\bar{x} - x_{(1)}}$$

**Remarks 1.** Whenever the given probability function involves a constant and the range of the variable is dependent on the parameter(s) to be estimated, first of all we should determine the constant by taking the total probability as unity and then proceed with the estimation part.

2. From the last two examples, it is obvious that whenever the range of the variable involves the parameter(s) to be estimated, the likelihood equations fail to give us valid estimates and in this case M.L.Es are obtained by adopting some other approach of maximising  $L$  or  $\log L$  directly.

**Example 5.39.** Obtain the maximum likelihood estimate of  $\theta$  in

$$f(x, \theta) = (1 + \theta)x^\theta, 0 < x < 1,$$

based on an independent sample of size  $n$ . Examine whether this estimate is sufficient for  $\theta$ .

**Solution.**

$$L(x, \theta) = \prod_{i=1}^n f(x_i, \theta) = (1 + \theta)^n \cdot \left( \prod_{i=1}^n x_i \right)^\theta$$

$$\Rightarrow \log L = n \log(1 + \theta) + \theta \cdot \sum_{i=1}^n \log x_i$$

$$\frac{\partial}{\partial \theta} \log L = \frac{n}{1 + \theta} + \sum_{i=1}^n \log x_i = 0$$

$$\Rightarrow n + \theta \sum_i \log x_i + \sum_i \log x_i = 0$$

$$\therefore \hat{\theta} = \frac{-n}{\sum_{i=1}^n \log x_i} - 1 = \frac{-n}{\log \left( \prod_{i=1}^n x_i \right)} - 1 \quad \dots (*)$$

$$\text{Also } L(x, \theta) = \left\{ (1 + \theta)^n \cdot \left( \prod_{i=1}^n x_i \right)^{\theta - 1} \right\} \cdot \left( \prod_{i=1}^n x_i \right)$$

Hence by Factorisation theorem,  $T = \left( \prod_{i=1}^n x_i \right)$  is a sufficient statistic for  $\theta$ , and  $\hat{\theta}$  being a one to one function of sufficient statistic  $\left( \prod_{i=1}^n x_i \right)$ , is also sufficient for  $\theta$ .

**Example 15-40.** (a) Obtain the most general form of distribution differentiable in  $\theta$ , for which the sample mean is the M.L.E.

[Delhi Univ. B.Sc. (Stat. Hons.), 1983]

(b) Show that the most general continuous distribution for which the M.L.E. of a parameter  $\theta$  is the geometric mean of the sample is

$$f(x, \theta) = \left( \frac{x}{\theta} \right)^{\theta \cdot \frac{\partial \psi}{\partial \theta}} \exp [\psi(\theta) + \xi(x)],$$

where  $\psi(\theta)$  and  $\xi(x)$  are arbitrary functions of  $\theta$  and  $x$  respectively.

**Solution.** (a) We have  $L = \prod_{i=1}^n f(x_i, \theta)$

$$\Rightarrow \log L = \sum_{i=1}^n \log f(x_i, \theta) = \sum_x \log f, \quad [f = f(x, \theta)]$$

the summation extending to all the values of  $x = (x_1, x_2, \dots, x_n)$  in the sample. The likelihood equation is

$$\begin{aligned} \frac{\partial}{\partial \theta} \log L &= 0, \text{ i.e., } \frac{\partial}{\partial \theta} \left( \sum_x \log f \right) = 0 \\ \Rightarrow \sum_x \frac{\partial}{\partial \theta} \log f &= 0 \quad \Rightarrow \quad \sum_x \frac{1}{f} \cdot \frac{\partial f}{\partial \theta} = 0 \end{aligned} \quad \dots (*)$$

We are given that the solution of (\*) is

$$\begin{aligned} \theta &= \frac{1}{n} \sum x \quad \text{or} \quad n\theta = \sum x \\ \Rightarrow \sum_x (x - \theta) &= 0 \end{aligned} \quad \dots (**)$$

Since this is true for all values of  $x$  and  $\theta$ , we get from (\*) and (\*\*),

$$\frac{1}{f} \cdot \frac{\partial f}{\partial \theta} = A(x - \theta),$$

where  $A$  is independent of  $x$  but may be function of  $\theta$ . Let us take

$$A = \frac{\partial^2 \psi}{\partial \theta^2} \text{ where } \psi = \psi(\theta) \text{ is any arbitrary function of } \theta.$$

$$\text{Thus } \frac{\partial}{\partial \theta} \log f = \frac{\partial^2 \psi}{\partial \theta^2} (x - \theta)$$

Integrating w.r. to  $\theta$  (partially), we get

$$\log f = (x - \theta) \cdot \frac{\partial \psi}{\partial \theta} - \int \frac{\partial \psi}{\partial \theta} (-1) d\theta + \xi(x) + k$$

where  $\xi(x)$  is an arbitrary function of  $x$  and  $k$  is arbitrary constant.

$$\therefore \log f = (x - \theta) \cdot \frac{\partial \psi}{\partial \theta} + \psi(\theta) + \xi(x) + k$$

$$\text{Hence } f = \text{Const. exp} \left[ (x - \theta) \cdot \frac{\partial \psi}{\partial \theta} + \psi(\theta) + \xi(x) \right]$$

which is the probability function of the required distribution.

**Remark.** In particular, if we take.

$$\psi(\theta) = \frac{\theta^2}{2} \text{ and } \xi(x) = -\frac{x^2}{2}, \text{ then}$$

$$\begin{aligned} f &= \text{Const. exp} \left[ (x - \theta) \cdot \theta + \frac{\theta^2}{2} - \frac{x^2}{2} \right] \\ &= \text{Const. exp} \left[ -\frac{1}{2}(x^2 + \theta^2 - 2\theta x) \right] \\ &= \text{Const. exp} \left\{ -\frac{1}{2}(x - \theta)^2 \right\} \end{aligned}$$

which is the probability function of the normal distribution with mean  $\theta$  and unit variance.

(b) Here the solution of the likelihood equation

$$\frac{\partial}{\partial \theta} \log L = \sum_x \frac{\partial}{\partial \theta} \log f = 0 \quad \dots(*)$$

$$\text{is } \theta = (x_1, x_2, \dots, x_n)^{1/n}$$

$$\Rightarrow \log \theta = \frac{1}{n} \sum_x \log x \Rightarrow \sum_x (\log x - \log \theta) = 0 \quad \dots(**)$$

Since this is true for all  $x$  and all  $\theta$ , we get from (\*) and (\*\*)

$$\frac{\partial}{\partial \theta} \log f = (\log x - \log \theta) A(\theta)$$

where  $A(\theta)$  is an arbitrary function of  $\theta$  and is independent of  $x$ .

Integrating w.r. to  $\theta$  (partially), we get

$$\log f = \log x \int A(\theta) d\theta - \int A(\theta) \log \theta d\theta + \xi(x)$$

where  $\xi(x)$  is an arbitrary function of  $x$  alone.

If we take  $\int A(\theta) d\theta = A_1(\theta)$ , then

$$\begin{aligned} \log f &= \log x \cdot A_1(\theta) - \left[ A_1(\theta) \log \theta - \int A_1(\theta) \cdot \frac{1}{\theta} d\theta \right] + \xi(x) \\ &= A_1(\theta) \log(x/\theta) + \int \frac{A_1(\theta)}{\theta} d\theta + \xi(x) \end{aligned}$$

Let us take

$$A_1(\theta) = \theta \frac{\partial \psi}{\partial \theta}, \text{ (suggested by the answer)}$$

where  $\psi = \psi(\theta)$  is an arbitrary function of  $\theta$  alone.

$$\therefore \log f = \theta \frac{\partial \psi}{\partial \theta} \log(x/\theta) + \int \frac{\partial \psi}{\partial \theta} d\theta + \xi(x)$$

$$= \theta \frac{\partial \psi}{\partial \theta} \cdot \log(x/\theta) + \psi(\theta) + \xi(x)$$

$$= \log \left[ \left( \frac{x}{\theta} \right)^{\theta} \frac{\partial \psi}{\partial \theta} \right] + \psi(\theta) + \xi(x)$$

Hence  $f = f(x, \theta) = \left( \frac{x}{\theta} \right)^{\theta} \frac{\partial \psi}{\partial \theta} \cdot \exp [\psi(\theta) + \xi(x)].$

**Example 15.41.** A sample of size  $n$  is drawn from each of the four normal populations which have the same variance  $\sigma^2$ . The means of the four populations are  $a+b+c$ ,  $a+b-c$ ,  $a-b+c$  and  $a-b-c$ . What are the M.L.Es. for  $a$ ,  $b$ ,  $c$ , and  $\sigma^2$ ?

**Solution.** Let the sample observations be denoted by  $x_{ij}$ ,  $i = 1, 2, 3, 4$ ;  $j = 1, 2, \dots, n$ . Since the four samples, from the four normal populations are independent, the likelihood function  $L$  of all the sample observations  $x_{ij}$ , ( $i = 1, 2, 3, 4$ ;  $j = 1, 2, \dots, n$ ), is given by

$$L = \left( \frac{1}{\sqrt{2\pi} \sigma} \right)^{4n} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^4 \sum_{j=1}^n (x_{ij} - \mu_i)^2 \right\}$$

where  $\mu_i$ , ( $i = 1, 2, 3, 4$ ) is mean of the  $i$ th population.

$$\therefore L = \left( \frac{1}{\sqrt{2\pi} \sigma} \right)^{4n} \cdot \exp \left[ -\frac{1}{2\sigma^2} \left\{ \sum_j (x_{1j} - \mu_1)^2 + \sum_j (x_{2j} - \mu_2)^2 + \sum_j (x_{3j} - \mu_3)^2 + \sum_j (x_{4j} - \mu_4)^2 \right\} \right]$$

$$\therefore \log L = k - 2n \log \sigma^2$$

$$-\frac{1}{2\sigma^2} \left[ \sum_j (x_{1j} - a - b - c)^2 + \sum_j (x_{2j} - a - b + c)^2 + \sum_j (x_{3j} - a + b - c)^2 + \sum_j (x_{4j} - a + b + c)^2 \right]$$

where  $k$  is a constant w.r. to  $a$ ,  $b$ ,  $c$  and  $\sigma^2$ .

The M.L.Es. for  $a$ ,  $b$ ,  $c$  and  $\sigma^2$  are the solutions of the simultaneous equations (maximum likelihood equations for estimating  $a$ ,  $b$ ,  $c$  and  $\sigma^2$ ):

$$\frac{\partial}{\partial a} \log L = 0 \quad \dots(1) \quad \frac{\partial}{\partial b} \log L = 0 \quad \dots(2)$$

$$\frac{\partial}{\partial c} \log L = 0 \quad \dots(3) \quad \frac{\partial}{\partial \sigma^2} \log L = 0 \quad \dots(4)$$

(1) gives

$$-\frac{1}{2\sigma^2} \left[ \sum_j (x_{1j} - a - b - c)(-2) + \sum_j (x_{2j} - a - b + c)(-2) + \sum_j (x_{3j} - a + b - c)(-2) + \sum_j (x_{4j} - a + b + c)(-2) \right] = 0$$

$$\begin{aligned} \Rightarrow & \sum_j (x_{1j} + x_{2j} + x_{3j} + x_{4j}) \\ & + n [((-a-b-c) + (-a-b+c) + (-a+b-c) + (-a+b+c))] = 0 \\ \Rightarrow & \sum_{j=1}^n \left( \sum_{i=1}^4 x_{ij} \right) + n(-4a) = 0. \\ \therefore & \hat{a} = \frac{1}{4n} \sum_{i=1}^4 \sum_{j=1}^n x_{ij} = \bar{x} \end{aligned}$$

Now (2) gives

$$\begin{aligned} -\frac{1}{2\sigma^2} & \left[ \sum_j (x_{1j} - a - b - c)(-2) + \sum_j (x_{2j} - a - b + c)(-2) \right. \\ & \quad \left. + \sum_j (x_{3j} - a + b - c)(2) + \sum_j (x_{4j} - a + b + c)(2) \right] = 0 \\ \Rightarrow & \sum_j x_{1j} + \sum_j x_{2j} - \sum_j x_{3j} - \sum_j x_{4j} \\ & + n[(-a-b-c) + (-a-b+c) - (-a+b-c) - (-a+b+c)] = 0 \\ \Rightarrow & \sum x_{1j} + \sum x_{2j} - \sum x_{3j} - \sum x_{4j} - 4nb = 0 \\ \therefore & \hat{b} = \frac{1}{4} \left[ \frac{1}{n} \sum x_{1j} + \frac{1}{n} \sum x_{2j} - \frac{1}{n} \sum x_{3j} - \frac{1}{n} \sum x_{4j} \right] \\ \Rightarrow & \hat{b} = (\bar{x}_1 + \bar{x}_2 - \bar{x}_3 - \bar{x}_4)/4, \end{aligned}$$

where  $\bar{x}_i$  is the mean of the  $i$ th sample.

Similarly (3) will give

$$\hat{c} = \frac{1}{4} (\bar{x}_1 - \bar{x}_2 + \bar{x}_3 - \bar{x}_4)/4$$

Equation (4) gives

$$\begin{aligned} -\frac{2n}{\sigma^2} + \frac{1}{2\sigma^4} & \left[ \sum_j (x_{1j} - a - b - c)^2 + \sum_j (x_{2j} - a - b + c)^2 \right. \\ & \quad \left. + \sum_j (x_{3j} - a + b - c)^2 + \sum_j (x_{4j} - a + b + c)^2 \right] = 0 \\ \therefore & \hat{\sigma}^2 = \frac{1}{4n} \left[ \sum_j (x_{1j} - \hat{a} - \hat{b} - \hat{c})^2 + \sum_j (x_{2j} - \hat{a} - \hat{b} + \hat{c})^2 \right. \\ & \quad \left. + \sum_j (x_{3j} - \hat{a} + \hat{b} - \hat{c})^2 + \sum_j (x_{4j} - \hat{a} + \hat{b} + \hat{c})^2 \right] \end{aligned}$$

**Example 15-42.** The following table gives probabilities and observed frequencies in four classes AB, Ab, aB and ab in a genetical experiment. Estimate the parameter  $\theta$  by the method of maximum likelihood and find its standard error.

Class	Probability	Observed frequency
$AB$	$\frac{1}{4}(2 + \theta)$	108
$A\bar{b}$	$\frac{1}{4}(1 - \theta)$	27
$aB$	$\frac{1}{4}(1 - \theta)$	30
$a\bar{b}$	$\frac{1}{4}\theta$	8

**Solution.** Using multinomial probability law, we have

$$L = L(\theta) = \frac{n!}{n_1! n_2! n_3! n_4!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}, \sum p_i = 1, \sum n_i = n$$

$$\Rightarrow \log L = C + n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 \log p_4,$$

where  $C = \log \left[ \frac{n!}{n_1! n_2! n_3! n_4!} \right]$ , is a constant.

$$\therefore \log L = C + n_1 \log \left( \frac{2+\theta}{4} \right) + n_2 \log \left( \frac{1-\theta}{4} \right) + n_3 \log \left( \frac{1-\theta}{4} \right) + n_4 \log \left( \frac{\theta}{4} \right)$$

Likelihood equation gives :

$$\frac{\partial \log L}{\partial \theta} = \frac{n_1}{2+\theta} - \frac{n_2}{1-\theta} - \frac{n_3}{1-\theta} + \frac{n_4}{\theta} = 0 \quad \dots (*)$$

$$\Rightarrow \frac{n_1}{2+\theta} - \frac{(n_2+n_3)}{1-\theta} + \frac{n_4}{\theta} = 0$$

Taking  $n_1 = 108, n_2 = 27, n_3 = 30$  and  $n_4 = 8$ , we get

$$\frac{108}{2+\theta} - \frac{(27+30)}{1-\theta} + \frac{8}{\theta} = 0$$

$$\Rightarrow 108\theta(1-\theta) - 57\theta(2+\theta) + 8(1-\theta)(2+\theta) = 0$$

$$\Rightarrow 173\theta^2 + 14\theta - 16 = 0$$

$$\Rightarrow \theta = \frac{-14 \pm \sqrt{196 + 11072}}{346} = -0.34 \text{ and } 0.26$$

But  $\theta$ , being the probability cannot be negative. Hence M.L.E. of  $\theta$  is given by  $\hat{\theta} = 0.26$   $\dots (**)$

Differentiating (\*) again partially w.r. to  $\theta$ , we get

$$\frac{\partial^2 \log L}{\partial \theta^2} = \frac{-n_1}{(2+\theta)^2} - \frac{(n_2+n_3)}{(1-\theta)^2} - \frac{n_4}{\theta^2}$$

$$- E \left( \frac{\partial^2 \log L}{\partial \theta^2} \right) = \frac{E(n_1)}{(2+\theta)^2} + \frac{E(n_2) + E(n_3)}{(1-\theta)^2} + \frac{E(n_4)}{\theta^2}$$

$$= \frac{np_1}{(2+\theta)^2} + \frac{n(p_2 + p_3)}{(1-\theta)^2} + \frac{np_4}{\theta^2}$$

$$= \frac{n(2+\theta)}{4(2+\theta)^2} + \frac{n(1-\theta)}{2(1-\theta)^2} + \frac{n\theta}{4\theta^2}$$

$$\therefore I(\theta) = \frac{n}{4(2+\theta)} + \frac{n}{2(1-\theta)} + \frac{n}{4\theta}; n = \sum n_i = 173.$$

$$\begin{aligned}
 &= 173 \left[ \frac{1}{4 \times 2.26} + \frac{1}{2 \times 0.74} + \frac{1}{4 \times 0.26} \right] \\
 &= 173 [0.11 + 0.67 + 0.96] = 173 \times 1.74 = 301.02 \\
 \text{S.E.}(\hat{\theta}) &= \sqrt{1/I(\theta)} = \frac{1}{\sqrt{301.02}} = 0.0576
 \end{aligned}$$

[c.f. (15-55) Theorem 15-13)]

**15-12. Method of Minimum Variance.** [Minimum Variance Unbiased Estimates (M.V.U.E.)]. In this section we shall look for estimates which (i) are unbiased and (ii) have minimum variance.

If  $L = \prod_{i=1}^n f(x_i, \theta)$ , is the likelihood function of a random sample of  $n$  observations  $x_1, x_2, \dots, x_n$  from a population with probability function  $f(x, \theta)$ , then the problem is to find a statistic  $t = t(x_1, x_2, \dots, x_n)$ , such that

$$E(t) = \int_{-\infty}^{\infty} t \cdot L \, dx = \gamma(\theta) \Rightarrow \int_{-\infty}^{\infty} [t - \gamma(\theta)] L \, dx = 0 \quad \dots(15-57)$$

$$\text{and } V(t) = \int_{-\infty}^{\infty} [t - E(t)]^2 L \, dx = \int_{-\infty}^{\infty} [t - \gamma(\theta)]^2 L \, dx \quad \dots(15-58)$$

is minimum, where

$\int_{-\infty}^{\infty} dx$  represents the  $n$ -fold integration

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} dx_1 dx_2 \cdots dx_n$$

In other words, we have to minimise (15-58) subject to the condition (15-57).

For detailed discussion of this method see MVU Estimators (§ 15-5-2) and Cramer-Rao Inequality (§ 15-7).

**15-13. Method of Moments.** This method was discovered and studied in detail by Karl Pearson.

Let  $f(x; \theta_1, \theta_2, \dots, \theta_k)$  be the density function of the parent population with  $k$  parameters  $\theta_1, \theta_2, \dots, \theta_k$ . If  $\mu'_r$  denotes the  $r$ th moment about origin, then

$$\mu'_r = \int_{-\infty}^{\infty} x^r f(x; \theta_1, \theta_2, \dots, \theta_k) dx, \quad (r = 1, 2, \dots, k) \quad \dots(15-59)$$

In general  $\mu'_1, \mu'_2, \dots, \mu'_k$  will be functions of the parameters  $\theta_1, \theta_2, \dots, \theta_k$ .

Let  $x_i, i = 1, 2, \dots, n$  be a random sample of size  $n$  from the given population. The method of moments consists in solving the  $k$ -equations (15-59) for  $\theta_1, \theta_2, \dots, \theta_k$  in terms of  $\mu'_1, \mu'_2, \dots, \mu'_k$  and then replacing these moments  $\mu'_r; r = 1, 2, \dots, k$  by the sample moments.

$$\text{e.g., } \hat{\theta}_i = \theta_i(\hat{\mu}_1', \hat{\mu}_2', \dots, \hat{\mu}_k') \\ = \theta_i(m_1', m_2', \dots, m_k'); i = 1, 2, \dots, k$$

where  $m_i$  is the  $i$ th moment about origin in the sample.

Then by the method of moments  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$  are the required estimators of  $\theta_1, \theta_2, \dots, \theta_k$  respectively.

**Remarks.** 1. Let  $(x_1, x_2, \dots, x_n)$  be a random sample of size  $n$  from a population with p.d.f.  $f(x, \theta)$ . Then  $X_i$ , ( $i = 1, 2, \dots, n$ ) are i.i.d.  $\Rightarrow X_i'$ , ( $i = 1, 2, \dots, n$ ) are i.i.d. r.v.'s. Hence if  $E(X_i')$  exists, then by W.L.L.N., we get

$$\frac{1}{n} \sum_{i=1}^n x_i' \xrightarrow{P} E(X_1') \\ \Rightarrow m_i' \xrightarrow{P} \mu_i' \quad \dots(15-60)$$

Hence the sample moments are consistent estimators of the corresponding population moments.

2. It has been shown that under quite general conditions, the estimates obtained by the method of moments are asymptotically normal but not, in general, efficient.

3. Generally the method of moments yields less efficient estimators than those obtained from the principle of maximum likelihood. The estimators obtained by the method of moments are identical with those given by the method of maximum likelihood if the probability mass function or probability density function is of the form

$$f(x, \theta) = \exp [b_0 + b_1 x + b_2 x^2 + \dots] \quad \dots(15-61)$$

where  $b$ 's are independent of  $x$  but may depend on  $\theta = (\theta_1, \theta_2, \dots)$ .

(15-61) implies that

$$L(x_1, x_2, \dots, x_n; \theta) = \exp [nb_0 + b_1 \sum x_i + b_2 \sum x_i^2 + \dots] \\ \Rightarrow \frac{\partial}{\partial \theta_j} \log L = a_0 + a_1 \sum x_i + a_2 \sum x_i^2 + a_3 \sum x_i^3 + \dots \quad \dots(15-61a)$$

Thus both the methods yield identical estimators if MLE's are obtained as linear functions of the moments.

**Example 15-43.** Estimate  $\alpha$  and  $\beta$  in the case of Pearson's Type III distribution by the method of moments.

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, 0 \leq x < \infty$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1988]

**Solution.** We have

$$\mu_r' = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^r x^{\alpha-1} e^{-\beta x} dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+r)}{\beta^{\alpha+r}} = \frac{\Gamma(\alpha+r)}{\Gamma(\alpha) \beta^r}$$

$$\mu_1' = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha) \beta} = \frac{\alpha}{\beta}, \quad \mu_2' = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha) \beta^2} = \frac{(\alpha+1)\alpha}{\beta^2}$$

$$\frac{\mu_2'}{\mu_1'^2} = \frac{\alpha+1}{\alpha} = \frac{1}{\alpha} + 1$$

$$\Rightarrow \alpha = \frac{\mu_1^2}{\mu_2' - \mu_1'^2}, \quad \beta = \frac{\alpha}{\mu_1'} = \frac{\mu_1}{\mu_2' - \mu_1'^2}$$

Hence  $\hat{\alpha} = \frac{m_1^2}{m_2' - m_1'^2}$  and  $\hat{\beta} = \frac{m_1'}{m_2' - m_1'^2}$

where  $m_1'$  and  $m_2'$  are the sample moments.

**Example 15-44.** For the double Poisson distribution :

$$p(x) = P(X=x) = \frac{1}{2} \cdot \frac{e^{-m_1} m_1^x}{x!} + \frac{1}{2} \cdot \frac{e^{-m_2} m_2^x}{x!}; x=0, 1, 2, \dots$$

show that the estimates for  $m_1$  and  $m_2$  by the method of moments are :

$$\mu_1' \pm \sqrt{\mu_2' - \mu_1' - \mu_1'^2}$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1993]

**Solution.** We have

$$\begin{aligned} \mu_1' &= \sum_{x=0}^{\infty} x \cdot p(x) = \frac{1}{2} \sum_{x=0}^{\infty} x \cdot \frac{e^{-m_1} m_1^x}{x!} + \frac{1}{2} \sum_{x=0}^{\infty} x \cdot \frac{e^{-m_2} m_2^x}{x!} \\ &= \frac{1}{2} m_1 + \frac{1}{2} m_2 \end{aligned} \quad \dots(*)$$

(since the first and second summations are the means of Poisson distributions with parameters  $m_1$  and  $m_2$  respectively).

$$\begin{aligned} \mu_2' &= \sum_{x=0}^{\infty} x^2 \cdot p(x) \\ &= \frac{1}{2} \left[ \sum_{x=0}^{\infty} x^2 \cdot \left( \frac{e^{-m_1} m_1^x}{x!} \right) + \sum_{x=0}^{\infty} x^2 \cdot \left( \frac{e^{-m_2} m_2^x}{x!} \right) \right] \\ &= \frac{1}{2} [(m_1^2 + m_1) + (m_2^2 + m_2)] \quad [\text{c.f. } \S \text{ 7.3.3}] \end{aligned}$$

$$\begin{aligned} \mu_2' &= \frac{1}{2} [(m_1 + m_2) + (m_1^2 + m_2^2)] \quad \dots(**) \\ &= \frac{1}{2} [2\mu_1' + m_1^2 + (2\mu_1' - m_1)^2] \quad [\text{Using (*)}] \\ &= \frac{1}{2} [2\mu_1' + m_1^2 + 4\mu_1'^2 + m_1^2 - 4m_1\mu_1'] \end{aligned}$$

$$\mu_2' = \mu_1' + m_1^2 + 2\mu_1'^2 - 2\mu_1'm_1 \Rightarrow m_1^2 - 2m_1\mu_1' + (2\mu_1'^2 + \mu_1' - \mu_2') = 0$$

$$\Rightarrow \hat{m}_1 = \frac{2\mu_1' \pm \sqrt{4\mu_1'^2 - 4(2\mu_1'^2 + \mu_1' - \mu_2')}}{2} = \mu_1' \pm \sqrt{\mu_2' - \mu_1' - \mu_1'^2}$$

Similarly on substituting for  $m_1$  in terms of  $m_2$  from (\*) in (\*\*), we get

$$m_2^2 - 2m_2\mu_1' + (2\mu_1'^2 + \mu_1' - \mu_2') = 0$$

Solving for  $m_2$ , we will get

$$\hat{m}_2 = \mu_1' \pm \sqrt{\mu_2' - \mu_1' - \mu_1'^2}$$

**Example 15-45.** A random variable  $X$  takes the values 0, 1, 2, with respective probabilities

$$\frac{\theta}{4N} + \frac{1}{2} \left(1 - \frac{\theta}{N}\right) \cdot \frac{\theta}{2N} + \frac{\alpha}{2} \left(1 - \frac{\theta}{N}\right) \quad \text{and} \quad \frac{\theta}{4N} + \frac{1-\alpha}{2} \left(1 - \frac{\theta}{N}\right)$$

where  $N$  is a known number and  $\alpha, \theta$  are unknown parameters. If 75 independent observations on  $X$  yielded the values 0, 1, 2 with frequencies 27, 38, 10 respectively, estimate  $\theta$  and  $\alpha$  by the method of moments.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

**Solution.**

$$\begin{aligned} E(X) &= 0 \cdot \left[ \frac{\theta}{4N} + \frac{1}{2} \left(1 - \frac{\theta}{N}\right) \right] + 1 \cdot \left[ \frac{\theta}{N} + \frac{\alpha}{2} \left(1 - \frac{\theta}{N}\right) \right] \\ &\quad + 2 \left[ \frac{\theta}{4N} + \frac{1-\alpha}{2} \left(1 - \frac{\theta}{N}\right) \right] \\ &= \frac{\theta}{N} + \left(1 - \frac{\theta}{N}\right) \left[ \frac{\alpha}{2} + (1 - \alpha) \right] \end{aligned}$$

$$\Rightarrow \mu_1' = \frac{\theta}{N} + \left(1 - \frac{\theta}{N}\right) \left(1 - \frac{\alpha}{2}\right) \\ = 1 - \frac{\alpha}{2} \left(1 - \frac{\theta}{N}\right) \quad \dots(*)$$

$$\begin{aligned} E(X^2) &= 1^2 \cdot \left[ \frac{\theta}{2N} + \frac{\alpha}{2} \left(1 - \frac{\theta}{N}\right) \right] + 2^2 \cdot \left[ \frac{\theta}{4N} + \frac{1-\alpha}{2} \left(1 - \frac{\theta}{N}\right) \right] \\ &= \frac{3\theta}{2N} + \left(1 - \frac{\theta}{N}\right) \left[ \frac{\alpha}{2} + 2(1 - \alpha) \right] \\ &= \frac{3\theta}{2N} + \left(1 - \frac{\theta}{N}\right) \left(2 - \frac{3\alpha}{2}\right) \end{aligned}$$

$$\Rightarrow \mu_2' = 2 - \frac{\theta}{2N} - \frac{3}{2} \alpha \left(1 - \frac{\theta}{N}\right) \quad \dots(**)$$

The sample frequency distribution is :

$x$	0	1		2	
$f$	21	38		10	

$$\mu_1' = \frac{1}{N} \sum f x = \frac{1}{75} (38 + 20) = \frac{58}{75}$$

$$\mu_2' = \frac{1}{N} \sum f x^2 = \frac{1}{75} (38 + 40) = \frac{78}{75}$$

Equating the sample moments to theoretical moments, we get

$$1 - \frac{\alpha}{2} \left(1 - \frac{\theta}{N}\right) = \frac{58}{75}$$

$$\Rightarrow \frac{\alpha}{2} \left(1 - \frac{\theta}{N}\right) = 1 - \frac{58}{75} = \frac{17}{75} \quad \dots(***)$$

Substituting in (\*\*), we get

$$2 - \frac{\theta}{2N} - 3 \times \frac{17}{75} = \frac{78}{75} \Rightarrow \hat{\theta} = \frac{42}{75} N$$

Substituting in (\*\*\*) , we get

$$\frac{\alpha}{2} \left( 1 - \frac{42}{75} \right) = \frac{17}{75} \Rightarrow \hat{\alpha} = \frac{34}{33}$$

**15.14. Method of Least Squares.**\* The principle of least squares is used to fit a curve of the form-

$$y = f(x, a_0, a_1, \dots, a_n) \quad \dots(15.62)$$

where  $a_i$ 's are unknown parameters, to a set of  $n$  sample observations  $(x_i, y_i)$ ;  $i = 1, 2, \dots, n$  from a bivariate population. It consists in minimising the sum of squares of residuals, viz.,

$$E = \sum_{i=1}^n [y_i - f(x_i, a_0, a_1, \dots, a_n)]^2 \quad \dots(15.63)$$

subject to variations in  $a_0, a_1, \dots, a_n$ .

The normal equations for estimating  $a_0, a_1, \dots, a_n$  are given by

$$\frac{\partial E}{\partial a_i} = 0; \quad i = 1, 2, \dots, n \quad \dots(15.64)$$

**Remarks.** 1. In chapter 9, we have discussed in detail the method of least squares for fitting linear regression (§ 9.1.1), polynomial regression (§ 9.1.3) and the exponential family of curves reducible to linear regression (§ 9.3). In chapter 10 § 10.12.1, we have discussed the method of fitting multiple linear regression.

2. If we are estimating  $f(x, a_0, a_1, \dots, a_n)$  as a linear function of the parameters  $a_0, a_1, \dots, a_n$ , the  $x$ 's being known given values, the least square estimators obtained as linear functions of the  $y$ 's will be MVU estimators.

### EXERCISE 15(b)

1. (a) State and explain the principle of maximum likelihood for estimation of population parameter.

(b) (i) Describe the M.L. method of estimation and discuss five of its optimal properties.

(ii) Examine a situation when M.L. method fails and explain how you tackle such situations.

(c) Define the likelihood function for a random sample drawn from (i) a discrete population, (ii) a continuous population.

Find the likelihood function for a random sample of size  $n$  from each of the following populations :

(a) Normal ( $\mu, \sigma^2$ ), (b) Binomial ( $n, p$ ), (c) Poisson ( $\mu$ ), (d) Uniform on  $(a, b)$ .  
[Calcutta Univ. B.Sc. (Maths. Hons.), 1991]

\* For detailed discussion see Chapter 9.

2. (a) A random variable  $X$  takes the values 0 and 1 with respective probabilities  $p$  and  $1 - p$ . Obtain on the basis of random sample of size  $n$ , the maximum likelihood estimator of  $p$ .

(b) Obtain the maximum likelihood estimator for the distribution having the probability mass function :

$$f(x, \theta) = \theta^x (1 - \theta)^{1-x}, x = 0, 1; 0 \leq \theta \leq 1$$

[Calcutta Univ. B.Sc. (Maths. Hons.), 1986]

(c) Obtain the maximum likelihood estimator of  $\theta$  in the following cases :

$$(i) f(x, \theta) = \frac{1}{\theta} \cdot \exp(-x/\theta); x \geq 0, \theta > 0$$

$$(ii) f(x, \theta) = {}^n C_x \theta^x (1 - \theta)^{n-x}; x = 0, 1, 2, \dots, n$$

3. Suppose that  $X$  has a distribution  $N(\mu, \sigma^2)$ , that is, the p.d.f of  $X$  is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Using M.L. estimation, determine  $\mu$  and  $\sigma^2$ . What conclusions do you draw on the nature of the result so obtained ?

4. (a) Explain the technique of the method of maximum likelihood and give a formula for the large sample standard error of the maximum-likelihood estimator.

(b) For the distribution with p.d.f.

$f(x, \theta) = \theta e^{-\theta x}$ , ( $x \geq 0 ; \theta > 0$ ), find the maximum likelihood estimators of  $\theta$  and  $E(X)$ , and obtain their large-sample standard errors.

(c)  $X$  is a random variable such that

$$\begin{aligned} P(X \leq x) &= 0, \text{ for } x < 0 \\ &= 1 - e^{-x\theta}, \text{ for } x \geq 0 \end{aligned}$$

Based on  $n$  independent observations on  $X$ , obtain the maximum likelihood estimator of  $E(X)$ .

5. (a) Let  $X_1, X_2, \dots, X_n$  be a random sample from the distribution with probability density function :

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}; 0 < x < \infty, 0 < \theta < \infty$$

Find the maximum likelihood estimator of  $\theta$ .

[Madras Univ. B.Sc. Sept., 1988]

(b) For the distribution :

$$dF(x) = \frac{1}{\theta^p \Gamma(p)} \exp(-x/\theta) x^{p-1}; 0 \leq x < \infty, p > 0, \theta > 0$$

where  $p$  is known, find out the maximum likelihood estimate of  $\theta$  on the basis of a random sample of size  $n$  from the distribution. Find the variance of the estimate.

6. (a) If  $x_i$  ( $i = 1, 2, \dots, n$ ) is an observed random sample from the distribution having p.d.f.

$$f_\lambda(x) = \frac{\lambda^{k+1} x^k \exp(-\lambda x)}{\Gamma(k+1)}, x > 0$$

where  $\lambda > 0$  and  $k$  is a known constant, show that the ML estimator  $\hat{\lambda}$  for  $\lambda$  is  $(k+1)/\bar{x}$ . Show that the corresponding estimator is biased but consistent and that its asymptotic distribution for large  $n$  is

$$N(\lambda, \lambda^2/[n(k+1)]).$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1986]

(b) Derive the MLE of the mean  $\frac{\alpha}{\alpha + 2}$  of the beta distribution :

$$f(x) = [B(\alpha, 2)]^{-1} x^{\alpha-1} (1-x), 0 < x \leq 1, \alpha > 0.$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

7. (a) From a sample of size  $n$  from the population of  $X$ , determine the maximum likelihood estimates of the parameters  $a$  and  $b$  of the probability density

$$f(x) = \text{Constant exp } [-(x-a)/b]; x \geq a, b > 0, -\infty < a < \infty$$

[Calcutta Univ. B.Sc. (Maths Hons.), 1991]

(b) Let  $X_1, X_2, \dots, X_n$  be a random sample from the distribution with p.d.f.

$$f(x; \theta_1, \theta_2) = \begin{cases} \frac{1}{\theta_2} e^{-(x-\theta_1)/\theta_2}, & x \geq \theta_1, -\infty < \theta_1 < \infty, \theta_2 > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Obtain the maximum likelihood estimators for  $\theta_1$  and  $\theta_2$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1992]

(c) Given a sample of  $n$  independent observations from the distribution with density :  $f(x, \theta_1, \theta_2) = \theta_2^{-1} \exp [-(x-\theta_1)/\theta_2]$ ,  $\theta_1 \leq x < \infty$

Find the maximum-likelihood estimator of  $\theta_2$  when  $\theta_1$  is known and the maximum likelihood estimator of  $\theta_1$  when  $\theta_2$  is known and also the joint maximum likelihood estimators of  $\theta_1$  and  $\theta_2$ . Comment on the estimators you obtain.

8. (a) A random variable  $X$  has the probability density function :

$$\begin{aligned} f(x) &= (\beta + 1) x^\beta, \text{ for } (0 < x < 1), (\beta > -1). \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

Based on  $n$ -independent observations on  $X$ , obtain the maximum likelihood estimator of  $\beta$  and an unbiased estimator of  $(\beta + 1)/(\beta + 2)$ , when  $\beta \neq -2$ .

(b) A random variable  $X$  has a distribution with density function

$$\begin{aligned} f(x) &= (\alpha + 1) x^\alpha, \text{ (0} \leq x \leq 1, \alpha > -1) \\ &= 0, \quad \text{otherwise} \end{aligned}$$

and a random sample of size 8 produces the data :

$$0.2, 0.4, 0.8, 0.5, 0.7, 0.9, 0.8, 0.9.$$

Find the maximum likelihood estimate of the unknown parameter  $\alpha$ , it being given that  $\ln(0.0145152) \approx -4.2326$  ( $\ln$  denotes natural logarithm).

[Burdwan Univ. B.Sc. (Hons.), 1989]

(c) Find the MLE of  $\theta$  for a random sample of size  $n$  from the distribution :

$$f(x, \theta) = (\theta + 1) x^\theta, \quad 0 \leq x \leq 1 \\ = 0, \quad \text{otherwise}$$

Show that it is also sufficient statistic for  $\theta$ .

Ans. MLE  $\hat{\theta} = \left[ -\frac{n}{\sum_{i=1}^n \log x_i} - 1 \right] \dots (*)$

$T = \prod_{i=1}^n x_i$ , is sufficient estimator for  $\theta$

$$\Rightarrow \hat{\theta} = \left[ \frac{-n}{\log(T/x_i)} - 1 \right], \text{ being a one to one function of}$$

sufficient statistic, is also a sufficient statistic for  $\theta$ .

9. (a) Obtain the MLE for the parameter  $\theta$  in a random sample of size  $n$  from the uniform population  $U[0, \theta]$ .

Ans.  $\hat{\theta} = x_{(n)}$ , the largest sample observation.

(b) Show by means of an example, that MLE are not, in general unique.

Ans. See Example 15-34.

(c) Show that in a random sample from a distribution with p.d.f.

$$f(x, \theta) = \theta e^{-\theta x}, x \geq 0$$

$1/\bar{X}$  is the MLE for  $\theta$  and has greater variance than the unbiased estimator  $(n-1)/(n\bar{X})$ .

Hint. MLE  $\hat{\theta} = \frac{1}{\bar{X}} = \frac{n}{T}$ ,  $T = \sum_{i=1}^n X_i \Rightarrow n\bar{X} = T$

$X_i$ , ( $i = 1, 2, \dots, n$ ) are i.i.d.  $\gamma(\theta, 1)$

$$\Rightarrow T = \sum_i X_i \sim \gamma(\theta, n)$$

$$E\left[\frac{n-1}{n\bar{X}}\right] = E\left[\frac{n-1}{T}\right] = (n-1)E(1/T) = \theta.$$

$$\text{Var}\left(\frac{n-1}{n\bar{X}}\right) = \left(\frac{n-1}{n}\right)^2 \text{Var}\left(\frac{1}{\bar{X}}\right) < \text{Var}\left(\frac{1}{\bar{X}}\right) = \text{Var}\hat{\theta}$$

10. (a) Let  $x_1, x_2, \dots, x_n$  be a random sample from a population with density :

$$f(x, \theta) = \frac{1}{2} \exp[-|x - \theta|], -\infty < x < \infty.$$

Find the estimator for  $\theta$  based on the method of maximum likelihood.

[Madras Univ. B.Sc., 1989]

Hint.  $L = \left(\frac{1}{2}\right)^n \exp\left[-\sum_{i=1}^n |x_i - \theta|\right]$  is maximum, if  $\sum_{i=1}^n |x_i - \theta|$  is minimum.  $\Rightarrow \hat{\theta} = \text{Median of } (x_1, x_2, \dots, x_n)$ .

(b) Obtain the maximum likelihood estimator of  $\theta$  based on a random sample of size  $n$  from the population with p.d.f.

$$(i) f(x, \theta) = e^{-(x-\theta)}; \theta \leq x < \infty, -\infty < \theta < \infty$$

$$(ii) f(x, \theta) = \theta x^{\theta-1}; 0 < x < 1, 0 < \theta < \infty.$$

Examine in each case, whether  $\theta$  is unbiased.

Hint. (i)  $L$  is maximum if  $\sum_{i=1}^n (x_i - \theta)$  is minimum.

$\Rightarrow$  Each deviation  $(x_i - \theta), i = 1, 2, \dots, n$  is minimum  $\Rightarrow \hat{\theta} = x_{(1)}$ .

11. (a) Explain what is meant by an estimate of a population parameter. Find the maximum likelihood estimate of the parameter  $\theta$  of a population having density function :

$$2(\theta - x)/\theta^2, (0 < x < \theta)$$

for a sample of unit size and examine whether the estimate so obtained is biased or not. [Calcutta Univ. B.Sc. (Maths. Hons.), 1987]

Ans.  $\hat{\theta} = 2x$ ; biased.

(b) Obtain Maximum Likelihood Estimator of  $\theta$  for the distribution :

$$f(x, \theta) = \frac{C_0 \theta^x}{x!}; x = 0, 1, 2, \dots; \theta > 0,$$

$C_0$  is a constant. Also write the Maximum Likelihood Estimator of  $3\theta^2 + 4\theta + 5$ . [Agra Univ. B.Sc., 1988]

Hint. For MLE of  $3\theta^2 + 4\theta + 5$ , use Invariance Property of MLE (c.f. Theorem 15.17)

(c) A population has a density function given by :

$$f(x) = 2v \sqrt{\frac{y}{\pi}} x^2 e^{-vx^2}; -\infty < x < \infty$$

Find the maximum likelihood estimate for  $v$ .

[Calcutta Univ. B.Sc. (Maths. Hons.), 1988]

12. (a) Consider a population made up of 3 different types of individuals occurring in the population with probabilities  $\theta^2$ ,  $2\theta(1-\theta)$  and  $(1-\theta)^2$ , respectively where  $0 < \theta < 1$ . Let  $n_1, n_2$  and  $n_3$  denote the respective random sample sizes of the above three types of individuals. Determine the maximum likelihood estimator for  $\theta$ . [Rajasthan PCS, 1989]

(b) Obtain the maximum likelihood estimate of  $\theta$ , if the variable takes the values 1, 2, 3 and 4 with probabilities  $(1-\theta)/2$ ,  $(1-\theta)/2$ ,  $\theta(1-\theta)$  and  $\theta^2$  respectively and the observed frequencies are  $n_1, n_2, n_3$  and  $n_4$  respectively.

13. In life-testing it is sometimes assumed that the life-time of an item is a random variable which is greater than or equal to  $x$  with probability

$$\exp\left[-\left(\frac{x}{\theta}\right)^m\right],$$

$x \geq 0, m > 0$  is known and  $\theta > 0$  is unknown. Suppose  $n$  such items are tested and yield  $X_1, X_2, \dots, X_n$  as their times of "death".

Find the maximum likelihood estimate of  $\theta$ .

14.  $X_1, X_2, X_3, X_4$  are independent normal random variables with means  $\alpha + \beta, \alpha - \beta, \alpha + 2\beta, \alpha - \beta$  respectively and a common variance  $\sigma^2$ , on the basis of one observation on each  $X_i$ ; obtain the maximum likelihood estimators of  $\alpha, \beta$  and  $\sigma^2$ . What is the asymptotic variance of  $\alpha^2$ ?

[Bharatiyan Univ. M.Sc. (Maths), 1993]

15. (a) For the bivariate normal distribution  $\Lambda(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$  find the maximum likelihood estimators

(i) of  $\sigma_1^2, \sigma_2^2$  and  $\rho$  when  $\mu_1$  and  $\mu_2$  are known,

(ii) of all five parameters of the distribution.

(b) Describe clearly the important properties to be possessed by a good estimator.

If  $(x_i, y_i)$ , ( $i = 1, 2, \dots, n$ ) come from a bivariate normal population with zero means, unit variances and co-efficient of correlation  $\rho$ , obtain the maximum likelihood estimator of  $\rho$ .

16. (a) Show that the most general continuous distribution for which the M.L.E. of a parameter  $\theta$  is the sample harmonic mean is :

$$f(x, \theta) = \exp \left[ \frac{1}{x} \left\{ \theta \frac{\partial \psi}{\partial \theta} - \psi(\theta) \right\} - \frac{\partial \psi}{\partial \theta} + \xi(x) \right]$$

where  $\psi(\theta)$  and  $\xi(x)$  are arbitrary functions of  $\theta$  and  $x$  respectively.

(b) Explain the principle of maximum likelihood estimation. Give examples to show that MLE need not be unique and also not necessarily unbiased.

Show that the most general form of the distribution for which the sample arithmetic mean  $\bar{X}$  is the MLE of  $\theta$  has the p.d.f.

$$f(x, \theta) = \exp [ (x - \theta) A'(\theta) + A(\theta) + B(x) ]$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

17. (a) Suppose that distribution of  $X$  is represented by the function :

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}; x = 0, 1, 2, \dots$$

where  $\lambda > 0$ . Given a random sample of size  $n$ , show that the sample mean is the maximum likelihood estimate of  $\lambda$ . Show further that this estimate is (i) best unbiased, and (ii) consistent. [Delhi Univ. M.A. (Eco.), 1986]

(b) Consider the estimation of the Poisson parameter from a random sample.

(i) Work out the maximum likelihood estimator and its variance.

(ii) Work out the Cramer – Rao Lower bound and show that it is equal to the variance worked out in (i). Comment on the significance of this result.

[Delhi Univ. M.A. (Eco.), 1990]

18.  $X$  is a discrete random variable and

$$P(X = r) = (1-p) p^{r-1}; r = 1, 2, 3, \dots$$

Find the MLE of  $p$  based on a random sample of  $n$  observations and its variance in large samples.

Show that the variance attains the lower bound of C.R. inequality.

19. Explain the terms : (i) sufficient estimator, (ii) efficient estimator, (iii) Cramer-Rao lower bound to the variance of an estimator, (iv) maximum likelihood estimator; and describe the relations amongst these four concepts.

20. (a) Describe the method of moments for estimating the parameters. What are the properties of the estimates obtained by this method ?

(b) Let  $(X_1, X_2, \dots, X_n)$  be a random sample from the p.d.f.

$$\begin{aligned} f(x, \theta) &= \theta e^{-\theta x}, 0 < x < \infty, \theta > 0 \\ &= 0, \quad \text{elsewhere} \end{aligned}$$

Estimate  $\theta$  using the method of moments.

(Madras Univ. B.Sc., 1988)

21.  $X_1, X_2, \dots, X_n$  is a random sample from

$$\begin{aligned} f(x; a, b) &= \frac{1}{b-a}; a < x < b \\ &= 0, \quad \text{elsewhere} \end{aligned}$$

Find estimates of  $a$  and  $b$  by the method of moments.

(Gujarat Univ. B.Sc. Oct., 1993)

22. Explain the methods of estimation-method of moments and maximum likelihood. Do these lead to the same estimates in respect of the standard deviation of a normal population ? Examine the properties of the estimates from the point of view of consistency and unbiasedness.

23. (a) Estimate  $\theta$  in the density function

$$f(x, \theta) = (1 + \theta) x^\theta; 0 < x < 1$$

by the method of moments and obtain the standard error of the estimator.

(b) The sample values from population with p.d.f.

$$f(x) = (1 + \theta) x^\theta, 0 < x < 1, \theta > 0,$$

are given below :

0.46, 0.38, 0.61, 0.82, 0.59, 0.53, 0.72, 0.44, 0.59, 0.60

Find the estimate of  $\theta$  by (i) method of moments and (ii) maximum likelihood estimation.

24. (a) For the distribution with probability function :

$$f(x, \theta) = \frac{e^{-\theta} \theta^x}{x! (1 - e^{-\theta})}; x = 1, 2, 3, \dots$$

obtain the estimate of  $\theta$  by the method of moments.

(b) For the following probability function :

$$f(x, p) = \binom{3}{x} \frac{p^x (1-p)^{3-x}}{1 - (1-p)^3}, [x = 1, 2, 3].$$

obtain the estimator of  $p$  by the method of moments, if the frequencies at  $x = 1, 2$  and 3 are respectively 22, 20 and 18.

-25. Let  $x_1, x_2, \dots, x_n$  be a sample from a distribution with density function :

$$f_\theta(x) = \theta(\theta + 1) x^{\theta-1} (1 - x), 0 < x < 1, \theta > 0$$

Determine the estimate of  $\theta$  by the method of moments.

[Indian Civil Services, 1981]

26. Explain the method of minimum chi-square in estimation, with a suitable example.

[Madras Univ. B.Sc., March 1989]

27. Describe the method of moments and discuss when the estimates obtained by the method of moments are identical with those of maximum likelihood estimates.

Estimate  $\alpha$  and  $\beta$  by the method of moments for the distribution :

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, 0 \leq x < \infty.$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1988]

28. State the conditions under which Maximum Likelihood Estimators of the parameters are identical with those given by the method of moments.

Examine if the MLEs of the parameter(s) are identical with those obtained by the method of moments in random sampling from the following distributions :

$$(i) f(x, \theta) = \frac{1}{\theta} \cdot \exp\left(-\frac{x}{\theta}\right); 0 < x < \infty$$

$$(ii) f(x, \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-(x - \mu)^2/2\sigma^2\right]; -\infty < x < \infty.$$

Ans. (i) MLE ( $\hat{\theta}$ ) =  $\bar{x} = \hat{\theta}$  (Method of Moments)

(ii) MLE ( $\hat{\mu}$ ) =  $\bar{X} = \hat{\mu}$  (Method of Moments)

MLE ( $\hat{\sigma}^2$ ) =  $s^2$  (sample variance) =  $\hat{\sigma}^2$  (Method of Moments).

29. Independent samples of sizes  $n_1$  and  $n_2$  are taken from two normal populations with equal means  $\mu$  and variances respectively equal to  $\lambda\sigma^2$ ,  $\sigma^2$ . Find the maximum likelihood estimator of  $\mu$  based on  $(n_1 + n_2)$  sample observations and show that its large sample variance is

$$\text{Var}(\hat{\mu}) = \sigma^2 / \left( \frac{n_1}{\lambda} + n_2 \right)$$

Hence show that the unbiased estimator,  $t = (n_1\bar{x}_1 + n_2\bar{x}_2) / (n_1 + n_2)$  has efficiency,  $\frac{\lambda(n_1 + n_2)^2}{(n_1\lambda + n_2)(n_1 + n_2\lambda)}$  which attains the value 1 if and only if  $\lambda = 1$ .

$$\text{Ans. MLE } (\hat{\mu}) = \left( \frac{n_1\bar{x}_1}{\lambda} + n_2\bar{x}_2 \right) / \left( \frac{n_1}{\lambda} + n_2 \right)$$

### OBJECTIVE TYPE QUESTIONS

1. Comment on the following statements :

- (i) In case of the Poisson distribution with parameter  $\lambda$ ,  $\bar{x}$  is sufficient for  $\lambda$ .
- (ii) If  $(X_1, X_2, \dots, X_n)$  be a sample of independent observations from the

uniform distribution on  $(\theta, \theta + 1)$ , then the maximum likelihood estimator of  $\theta$  is unique.

(iii) A maximum likelihood estimator is always unbiased.

(iv) Unbiased estimator is necessarily consistent

(v) A consistent estimator is also unbiased.

(vi) An unbiased estimator whose variance tends to zero as sample size increases is consistent.

(vii) If  $t$  is a sufficient statistic for  $\theta$  then  $f(t)$  is a sufficient statistic for  $f(\theta)$ .

(viii) If  $t_1$  and  $t_2$  are two independent estimators of  $\theta$ , then  $t_1 + t_2$  is less efficient than both  $t_1$  and  $t_2$ .

(ix) If  $T$  is a consistent estimator of a parameter  $\theta$ , then  $aT + b$  is a consistent estimator of  $a\theta + b$ , where  $a$  and  $b$  are constants.

(x) If  $x$  is the number of successes in  $n$  independent trials with a constant probability  $p$  of success in each trial, then  $x/n$  is a consistent estimator of  $p$ .

## II. Fill in the blanks :

(i) In a random sample of size  $n$  from a population with mean  $\mu$ , the sample mean ( $\bar{x}$ ) is ... estimate of ...

(ii) The sample median is ... estimate for the mean of normal population.

(iii) An estimator  $\hat{\theta}$  of a parameter  $\theta$  is said to be unbiased if ...

(iv) The variance  $s^2$  of a sample of size  $n$  is a ... estimator of population variance  $\sigma^2$ .

(v) If a sufficient estimator exists, it is a function of the ... estimator.

(vi) ... estimate may not be unique.

III. (a) Give example of a statistic  $t$  which is unbiased for a parameter  $\theta$  but  $t^2$  is not unbiased for  $\theta^2$ .

(b) Give example of an M.L. estimator which is not unbiased.

IV. What is the relationship between a sufficient estimator and a maximum likelihood estimator ?

V. (i) If  $\bar{x}$  is an unbiased estimator for the population mean  $\mu$ , state which of the following are unbiased estimators for  $\mu^2$ :

(a)  $\bar{x}^2$ , (b)  $\bar{x}^2 - \frac{\sigma^2}{n}$  ( $\sigma^2$  is known/unknown).

(ii) If  $t$  is the maximum likelihood estimator for  $\theta$ , state the condition under which  $f(t)$  will be the maximum likelihood estimator for  $f(\theta)$ .

(iii) Write down the condition for the Cramer-Rao lower bound for the variance of an unbiased estimator to be attained.

(iv) Write down the general form of the distribution admitting sufficient statistic.

VI. A random variable  $X$  takes the values 1, 2, 3 and 4, each with probability  $\frac{1}{4}$ . A random sample of three values of  $x$  is taken,  $\bar{x}$  is the mean and  $m$  is the median of this sample. Show that both  $\bar{x}$  and  $m$  are unbiased estimators

of the mean of the population, but  $\bar{x}$  is more efficient than  $m$ . Compare their efficiencies.

VII. Give an example of estimates which are

- (i) Unbiased and efficient, (ii) Unbiased and inefficient.

**15-15. Confidence Interval and Confidence Limits.** Let  $x_i$ , ( $i = 1, 2, \dots, n$ ) be a random sample of  $n$  observations from a population involving a single unknown parameter  $\theta$  (say). Let  $f(x, \theta)$  be the probability function of the parent distribution from which the sample is drawn and let us suppose that this distribution is continuous. Let  $t = t(x_1, x_2, \dots, x_n)$ , a function of the sample values be an estimate of the population parameter  $\theta$ , with the sampling distribution given by  $g(t, \theta)$ .

Having obtained the value of the statistic  $t$  from a given sample, the problem is, "Can we make some reasonable probability statements about the unknown parameter  $\theta$  in the population, from which the sample has been drawn?" This question is very well answered by the technique of *Confidence Interval* due to Neyman and is obtained below :

We choose once for all some small value of  $\alpha$  (5% or 1%) and then determine two constants say,  $c_1$  and  $c_2$  such that

$$P(c_1 < \theta < c_2 | t) = 1 - \alpha \quad \dots(15-65)$$

The quantities  $c_1$  and  $c_2$ , so determined, are known as the *confidence limits* or *fiducial limits* and the interval  $[c_1, c_2]$  within which the unknown value of the population parameter is expected to lie, is called the *confidence interval* and  $(1 - \alpha)$  is called the *confidence coefficient*.

Thus if we take  $\alpha = 0.05$  (or 0.01), we shall get 95% (or 99%) confidence limits.

**How to find  $c_1$  and  $c_2$ ?** Let  $T_1$  and  $T_2$  be two statistics such that

$$P(T_1 > \theta) = \alpha_1 \quad \dots(15-66)$$

$$\text{and} \quad P(T_2 < \theta) = \alpha_2 \quad \dots(15-66a)$$

where  $\alpha_1$  and  $\alpha_2$  are constants independent of  $\theta$ . (15-66) and (15-66a) can be combined to give

$$P(T_1 < \theta < T_2) = 1 - \alpha, \quad \dots(15-66b)$$

where  $\alpha = \alpha_1 + \alpha_2$ . Statistics  $T_1$  and  $T_2$  defined in (15-66) and (15-66a) may be taken as  $c_1$  and  $c_2$  defined in (15-65).

For example, if we take a large sample from a normal population with mean  $\mu$  and standard deviation  $\sigma$ , then

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$\text{and} \quad P(-1.96 < Z < 1.96) = 0.95,$$

[From Normal Probability Tables]

$$\Rightarrow P\left(-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

$$\Rightarrow P\left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$$

Thus  $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$  are 95% confidence limits for the unknown parameter  $\mu$ ,

the population mean and the interval

$\left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right]$  is called the 95% confidence interval.

Also

$$P(-2.58 < Z < 2.58) = 0.99$$

$$\Rightarrow P\left(-2.58 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 2.58\right) = 0.99$$

$$\Rightarrow P\left(\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}}\right) = 0.99$$

Hence 99% confidence limits for  $\mu$  are  $\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$  and 99% confidence interval for  $\mu$  is  $\left[\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}}\right]$ .

**Remarks 1.** Usually  $\sigma^2$  is not known and its unbiased estimate  $S^2$  obtained from the samples, is used. However if  $n$  is small,

$$Z = \frac{\bar{x} - \mu}{S/\sqrt{n}} \text{ is not } N(0, 1)$$

and in this case the confidence limits and confidence intervals for  $\mu$  are obtained by using Student's 't' distribution.

2. It can be seen that in many cases there exist more than one set of confidence intervals with the same confidence coefficient. Then the problem arises as to which particular set is to be regarded as better than the others in some useful sense and in such cases we look for the shortest of all the intervals.

**Example 15-45.** Obtain 100  $(1 - \alpha)\%$  confidence intervals for the parameters (a)  $\theta$  and (b)  $\sigma^2$ , of the normal distribution

$$f(x, \theta; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \theta}{\sigma}\right)^2\right], -\infty < x < \infty$$

**Solution.** Let  $X_i$ , ( $i = 1, 2, \dots, n$ ) be a random sample of size  $n$  from the density  $f(x; \theta, \sigma)$  and let

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

(a) The statistic :

$$t = \frac{\bar{X} - \theta}{S/\sqrt{n}}$$

follows student's  $t$ -distribution with  $(n - 1)$  degrees of freedom. Hence  $100(1 - \alpha)\%$  confidence limits for  $\theta$  are given by

$$\begin{aligned} P[|t| \leq t_\alpha] &= 1 - \alpha \\ \Rightarrow P\left[|\bar{X} - \theta| \leq \frac{S}{\sqrt{n}} t_\alpha\right] &= 1 - \alpha \\ \Rightarrow P\left[\bar{X} - t_\alpha \cdot \frac{S}{\sqrt{n}} \leq \theta \leq \bar{X} + t_\alpha \cdot \frac{S}{\sqrt{n}}\right] &= 1 - \alpha \quad \dots(567) \end{aligned}$$

where  $t_\alpha$  is the tabulated value of  $t$  for  $(n - 1)$  d.f. at significance level ' $\alpha$ '. Hence the required confidence interval for  $\theta$  is :

$$\left(\bar{X} - t_\alpha \frac{S}{\sqrt{n}}, \bar{X} + t_\alpha \frac{S}{\sqrt{n}}\right)$$

(b) Case (i)  $\theta$  is known and equal to  $\mu$  (say).

$$\text{Then } \frac{\sum(X_i - \mu)^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n)}$$

If we define  $\chi_\alpha^2$  as the value of  $\chi^2$  such that

$$P(\chi^2 > \chi_\alpha^2) = \int_{\chi_\alpha^2}^{\infty} p(\chi^2) d\chi^2 = \alpha \quad \dots(*)$$

where  $p(\chi^2)$  is the p.d.f. of  $\chi^2$ -distribution with  $n$  d.f., then the required confidence interval is given by

$$\begin{aligned} P[\chi^2_{1-(\alpha/2)} \leq \chi^2 \leq \chi^2_{\alpha/2}] &= 1 - \alpha \\ \Rightarrow P\left[\chi^2_{1-(\alpha/2)} \leq \frac{ns^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right] &= 1 - \alpha \quad \dots(**) \end{aligned}$$

$$\text{Now } \frac{ns^2}{\sigma^2} \leq \chi^2_{\alpha/2} \Rightarrow \frac{ns^2}{\chi^2_{\alpha/2}} \leq \sigma^2$$

$$\text{and } \chi^2_{1-(\alpha/2)} \leq \frac{ns^2}{\sigma^2} \Rightarrow \sigma^2 \leq \frac{ns^2}{\chi^2_{1-(\alpha/2)}}$$

Hence (\*\*) gives

$$P\left[\frac{ns^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{1-(\alpha/2)}}\right] = 1 - \alpha \quad \dots(***)$$

where  $\chi^2_{\alpha/2}$  and  $\chi^2_{1-(\alpha/2)}$  are obtained from (\*) by using  $n$  d.f.

Thus e.g., 95% confidence interval for  $\sigma^2$  is given by

$$P\left[\frac{ns^2}{\chi^2_{0.025}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{0.975}}\right] = 0.95$$

Case (ii).  $\theta$  is unknown. In this case the statistic

$$\frac{\sum(X_i - \bar{X})^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Here also confidence interval for  $\sigma^2$  is given by (\*\*\*\*) where now  $\chi^2_\alpha$  is the significant value of  $\chi^2$  [as defined in (\*)] for  $(n - 1)$  d.f. at the significance level ' $\alpha$ '.

**Example 15.46.** Show that the largest observations  $L$  of a sample of  $n$  observations from a rectangular distribution with density function :

$$\begin{aligned} f(x, \theta) &= \frac{1}{\theta}, \quad 0 \leq x \leq \theta \\ &= 0, \text{ otherwise} \end{aligned} \quad \dots (*)$$

has the distribution

$$dG(L) = n \left( \frac{L}{\theta} \right)^{n-1} \cdot \frac{dL}{\theta}, \quad 0 \leq L \leq \theta$$

Show that the distribution of  $V = L/\theta$  is given by p.d.f.

$$h(v) = nv^{n-1}, \quad 0 \leq v \leq 1$$

Hence deduce that the confidence limits for  $\theta$  corresponding to confidence coefficient  $\alpha$  are  $L$  and  $\frac{L}{(1-\alpha)^{1/n}}$

[Delhi Univ. B.Sc. (Stat. Hons.), 1982, 1983]

**Solution.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the population (\*) and let  $L = \max(X_1, X_2, \dots, X_n)$ . The distribution of  $L$  is given by  $dG(L) = n[F(L)]^{n-1} \cdot f(L) dL$ ,

where  $F(\cdot)$  is the distribution function of  $X$  given by

$$F(L) = \int_0^L f(x, \theta) dx = \frac{L}{\theta}$$

$$\therefore dG(L) = n \left( \frac{L}{\theta} \right)^{n-1} \cdot \frac{dL}{\theta}, \quad 0 \leq L \leq \theta$$

If we take  $V = L/\theta$ , the Jacobian of transformation is  $\theta$ . Hence p.d.f.  $h(\cdot)$  of  $V$  is given by

$$h(v) = nv^{n-1} \cdot \frac{1}{\theta} |J| = nv^{n-1}, \quad 0 \leq v \leq 1$$

which is independent of  $\theta$ .

To obtain the confidence limits for  $\theta$ , with confidence coefficient  $\alpha$ , let us define  $v_\alpha$  such that

$$P(v_\alpha < V < 1) = \alpha \Rightarrow \int_{v_\alpha}^1 h(v) dv = \alpha \quad \dots (**)$$

$$\Rightarrow n \int_{v_\alpha}^1 v^{n-1} dv = \alpha \Rightarrow 1 - v_\alpha^n = \alpha$$

$$\Rightarrow v_\alpha = (1 - \alpha)^{1/n} \quad \dots (***)$$

From (\*\*) and (\*\*\*), we get

$$P[(1 - \alpha)^{1/n} < V < 1] = \alpha$$

$$\Rightarrow P \left[ (1 - \alpha)^{1/n} < \frac{L}{\theta} < 1 \right] = \alpha$$

$$\Rightarrow P \left[ L < \theta < \frac{L}{(1 - \alpha)^{1/n}} \right] = \alpha$$

Hence the required confidence limits for  $\theta$  are  $L$  and  $L/(1 - \alpha)^{1/n}$ .

**Example 15.47.** Given a random sample from a population with p.d.f.

$$f(x, \theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta$$

show that  $100(1 - \alpha)\%$  confidence interval for  $\theta$  is given by  $R$  and  $R/\psi$  where  $\psi$  is given by

$$\psi^{n-1} [n - (n-1)\psi] = \alpha,$$

and  $R$  is the sample range.

**Solution.** The joint p.d.f. of  $x_1, x_2, \dots, x_n$  is given by

$$L = \frac{1}{\theta^n} \cdot 0 \leq x_i \leq \theta$$

If  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  is the ordered sample then the joint p.d.f. of  $x_{(n)}$  and  $x_{(1)}$  is given by

$$g(x_{(1)}, x_{(n)}) = \frac{n(n-1)}{\theta^n} [x_{(n)} - x_{(1)}]^{n-2}, \quad 0 \leq x_{(1)} \leq x_{(n)} \leq \theta$$

To obtain the distribution of the sample range  $R$ , let us make the transformation of variables

$$R = x_{(n)} - x_{(1)} \text{ and } v = x_{(1)} \Rightarrow v = x_{(n)} - R \leq \theta - R$$

The Jacobian of transformation is  $|J| = 1$  and the joint p.d.f. of  $R$  and  $V$  becomes

$$h(R, v) = \frac{n(n-1)}{\theta^n} R^{n-2}, \quad 0 < v < \theta - R$$

The marginal density of  $R$  is given by

$$\begin{aligned} h_1(R) &= \int_0^{\theta-R} \frac{n(n-1)}{\theta^n} \cdot R^{n-2} dv \\ &= \frac{n(n-1) R^{n-2} (\theta - R)}{\theta^n}, \quad 0 \leq R \leq \theta \end{aligned}$$

The density of  $U = R/\theta$  is

$$\begin{aligned} h_2(u) &= h_1(R) \left| \frac{dR}{du} \right| = \frac{n(n-1) R^{n-2} (\theta - R)}{\theta^n} \cdot \theta \\ &= n(n-1) u^{n-2} (1-u), \quad 0 \leq u \leq 1 \end{aligned}$$

$100(1 - \alpha)\%$  confidence interval for  $\theta$  is given by

$$P(\psi \leq U \leq 1) = 1 - \alpha \quad \dots (*)$$

where  $\psi$  is obtained from the equation

$$\int_0^\psi h_2(u) du = \alpha$$

$$\Rightarrow n(n-1) \int_0^\psi u^{n-2} (1-u) du = \alpha$$

$$\Rightarrow \left[ n u^{n-1} - (n-1) u^n \right]_0^\psi = \alpha$$

$$\Rightarrow \psi^{n-1} [n - (n-1)\psi] = \alpha \quad \dots (**)$$

From (\*), we get

$$\begin{aligned} P\left[\psi \leq \frac{R}{\theta} \leq 1\right] &= 1 - \alpha \\ \Rightarrow P\left[R \leq \theta \leq \frac{R}{\psi}\right] &= 1 - \alpha \end{aligned}$$

Hence the required limits for  $\theta$  are given by  $R$  and  $R/\psi$ , where  $\psi$  is the solution of (\*\*).

**Example 15.48.** Given one observation from a population with p.d.f.

$$f(x, \theta) = \frac{2}{\theta^2} (\theta - x), \quad 0 \leq x \leq \theta,$$

obtain  $100(1 - \alpha)\%$  confidence interval for  $\theta$ .

[Delhi Univ. B.Sc. (Stat Hons.), 1991]

**Solution.** The density of  $u = x/\theta$  is given by

$$\begin{aligned} g(u) &= f(x, \theta) \cdot \left| \frac{dx}{du} \right| = \frac{2}{\theta^2} (\theta - x) \cdot \theta \\ &= 2(1 - u), \quad 0 \leq u \leq 1 \end{aligned}$$

To obtain  $100(1 - \alpha)\%$  confidence interval for  $\theta$ , we choose two quantities  $u_1$  and  $u_2$  such that

$$P[u_1 \leq u \leq u_2] = 1 - \alpha \quad \dots(*)$$

$$\text{and} \quad P[u < u_1] = P[u > u_2] = \alpha/2$$

$$\text{Now} \quad P[u < u_1] = \frac{\alpha}{2} \Rightarrow \int_0^{u_1} 2(1 - u) du = \frac{\alpha}{2}$$

$$\Rightarrow u_1^2 - 2u_1 + \frac{\alpha}{2} = 0 \quad \dots(**)$$

$$\text{Similarly,} \quad P[u > u_2] = \frac{\alpha}{2} \Rightarrow \int_{u_2}^1 2(1 - u) du = \frac{\alpha}{2}$$

$$\Rightarrow u_2^2 - 2u_2 + \left(1 - \frac{\alpha}{2}\right) = 0 \quad \dots(***)$$

From (\*), we get

$$P\left[u_1 \leq \frac{x}{\theta} \leq u_2\right] = 1 - \alpha \Rightarrow P\left[\frac{x}{u_2} \leq \theta \leq \frac{x}{u_1}\right] = 1 - \alpha$$

Hence the required interval for  $\theta$  is  $\left(\frac{x}{u_2}, \frac{x}{u_1}\right)$ , where  $u_1$  and  $u_2$  are given by (\*\*) and (\*\*\*).

**15.15.1. Confidence Intervals for Large Samples.** It has been proved that under certain regularity conditions, the first derivative of the logarithm of the likelihood function w.r.t parameter  $\theta$  viz.,  $\frac{\partial}{\partial \theta} \log L$ , is asymptotically normal with mean zero and variance given by

$$\text{Var} \left( \frac{\partial}{\partial \theta} \log L \right) = E \left( \frac{\partial}{\partial \theta} \log L \right)^2 - E \left( - \frac{\partial^2}{\partial \theta^2} \log L \right)$$

Hence for large  $n$ ,

$$Z = \frac{\frac{\partial}{\partial \theta} \log L}{\sqrt{\text{Var} \left( \frac{\partial}{\partial \theta} \log L \right)}} \sim N(0, 1) \quad \dots(15-68)$$

The result enables us to obtain confidence interval for the parameter  $\theta$  in large samples. Thus for large samples, the confidence interval for  $\theta$  with confidence coefficient  $(1 - \alpha)$  is obtained by converting the inequalities in

$$P[|Z| \leq \lambda_\alpha] = 1 - \alpha \quad \dots(15-69)$$

where  $\lambda_\alpha$  is given by

$$\frac{1}{\sqrt{2\pi}} \int_{-\lambda_\alpha}^{\lambda_\alpha} \exp(-u^2/2) du = 1 - \alpha \quad \dots[15-69(a)]$$

**Example 15-49.** Obtain 100  $(1 - \alpha)\%$  confidence limits (for large samples) for the parameter  $\lambda$  of the Poisson distribution

$$f(x, \lambda) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}; x = 0, 1, 2, \dots$$

**Solution.** We have

$$\begin{aligned} \frac{\partial}{\partial \lambda} \log L &= \frac{\partial}{\partial \lambda} \left[ -n\lambda + \left( \sum_{i=1}^n x_i \right) \log \lambda - \sum_{i=1}^n \log x_i \right] \\ &= -n + \frac{\sum x_i}{\lambda} = n \left( \frac{\bar{x}}{\lambda} - 1 \right) \\ \text{Var} \left( \frac{\partial}{\partial \lambda} \log L \right) &= E \left( - \frac{\partial^2}{\partial \lambda^2} \log L \right) = E \left( \frac{n\bar{x}}{\lambda^2} \right) \\ &= \frac{n}{\lambda^2} E(\bar{x}) = \frac{n}{\lambda} \end{aligned}$$

$$\therefore Z = \frac{n \left( \frac{\bar{x}}{\lambda} - 1 \right)}{\sqrt{n/\lambda}} = \sqrt{(n/\lambda)} (\bar{x} - \lambda) \sim N(0, 1) \quad [\text{Using (15-68)}]$$

Hence 100  $(1 - \alpha)\%$  confidence interval for  $\lambda$  is given by (for large samples)

$$P[|\sqrt{(n/\lambda)} (\bar{x} - \lambda)| \leq \lambda_\alpha] = 1 - \alpha$$

Hence the required limits for  $\lambda$  are the roots of the equation :

$$|\sqrt{n/\lambda} (\bar{x} - \lambda)| = \lambda_\alpha$$

$$\Rightarrow n(\bar{x} - \lambda)^2 - \lambda \cdot \lambda_\alpha^2 = 0$$

$$\Rightarrow \lambda^2 - \lambda \left( 2\bar{x} + \frac{\lambda_{\alpha}^2}{n} \right) + \bar{x}^2 = 0$$

$$\Rightarrow \lambda = \frac{\left( 2\bar{x} + \frac{\lambda_{\alpha}^2}{n} \right) \pm \sqrt{\left[ \left( 2\bar{x} + \frac{\lambda_{\alpha}^2}{n} \right)^2 - 4\bar{x}^2 \right]^{1/2}}}{2} \quad \dots(*)$$

For example, 95% confidence interval for  $\lambda$  is given by taking  $\lambda_{\alpha} = 1.96$  in (\*), thus giving

$$\lambda = \frac{1}{2} \left( 2\bar{x} + \frac{3.84}{n} \right) \pm \sqrt{\left( \frac{3.84\bar{x}}{n} + \frac{3.69}{n^2} \right)} = \bar{x} \pm 1.96 \sqrt{\frac{\bar{x}}{n}},$$

to order  $n^{-1/2}$ .

**Example 15.50.** Show that for the distribution :

$$dF(x) = \theta e^{-x\theta}; 0 < x < \infty$$

central confidence limits for  $\theta$  for large samples with 95% confidence coefficient are given by

$$\theta = \left( 1 \pm \frac{1.96}{\sqrt{n}} \right) / \bar{x}$$

**Solution.** Here  $L = \theta^n \exp \left[ -\theta \sum_{i=1}^n x_i \right]$

$$\frac{\partial}{\partial \theta} \log L = \frac{\partial}{\partial \theta} [n \log \theta - \theta \sum x_i]$$

$$= \frac{n}{\theta} - \sum_{i=1}^n x_i = n \left( \frac{1}{\theta} - \bar{x} \right)$$

$$\frac{\partial^2}{\partial \theta^2} \log L = -\frac{n}{\theta^2}$$

$$\therefore \text{Var} \left( \frac{\partial}{\partial \theta} \log L \right) = E \left( -\frac{\partial^2}{\partial \theta^2} \log L \right) = \frac{n}{\theta^2}$$

Hence, for large samples, using (15.68) we have :

$$Z = \frac{n \left( \frac{1}{\theta} - \bar{x} \right)}{\sqrt{n/\theta^2}} \sim N(0, 1) \Rightarrow \sqrt{n} (1 - \theta \bar{x}) \sim N(0, 1)$$

Hence 95% confidence limits for  $\theta$  are given by

$$P[-1.96 \leq \sqrt{n} (1 - \theta \bar{x}) \leq 1.96] = 0.95$$

$$\text{Now } \sqrt{n} (1 - \theta \bar{x}) \leq 1.96 \Rightarrow \left( 1 - \frac{1.96}{\sqrt{n}} \right) \frac{1}{\bar{x}} \leq \theta \quad \dots(**)$$

$$\text{and } -1.96 \leq \sqrt{n} (1 - \theta \bar{x}) \Rightarrow \theta \leq \left( 1 + \frac{1.96}{\sqrt{n}} \right) \frac{1}{\bar{x}} \quad \dots(***)$$

Hence, from (\*) and (\*\*), the central 95% confidence limits for  $\theta$  are given by

$$\theta = \left( 1 \pm \frac{1.96}{\sqrt{n}} \right) \cdot \bar{x}$$

### EXERCISE 15 (c)

1. Discuss the concept of interval estimation and provide suitable illustration. [Delhi Univ. M.A. (Eco.), 1987]

2. Critically examine how interval estimation differs from point estimation. Give the 95% confidence interval for the mean of the normal distribution, when its variance is known. [Madras Univ. B.Sc. Sept., 1988]

3. What are confidence intervals ? How are they constructed using  $t$ -distribution ? [Madras Univ. B.Sc., March, 1989]

4. The random variable  $X$  is uniformly distributed in  $(a, a + 2)$ . Obtain limits  $x_1$  and  $x_2$  such that

$$P(X \leq x_1) = P(X \geq x_2) = 0.025$$

The random variable is observed once, the value being  $x_0$ . Give a method of obtaining an interval estimate for ' $a$ ' which you expect to be correct in 95% of trials. [Calcutta Univ. B.Sc. (Maths. Hons.), 1990]

5. Obtain 100  $(1 - \alpha)\%$  confidence interval either for the unknown parameter  $p$  of a binomial distribution when the parameter  $n$  is known or for the population correlation coefficient when the population is Normal.

[Delhi Univ. B.Sc. (Stat. Hons.), 1983]

6. Let  $f_\theta(x) = 1/\theta$ ,  $0 \leq x \leq \theta$  and let  $L$  be the largest observation of a sample of size  $n$  from the above distribution.

Obtain the distribution of  $(L/\theta)$  and hence deduce that the confidence limits corresponding to confidence coefficient  $\alpha$  are  $L$ , and  $\frac{L}{(1 - \alpha)^{1/n}}$  respectively.

[Delhi Univ. B.Sc. (Stat. Hons.), 1992]

7. (a) What are confidence intervals ?  $y$  is the largest observation in a sample of size  $n$  drawn from a rectangular population in  $(0, \theta)$ . Find the confidence coefficient corresponding to the confidence interval

$$\{y, y/(1 - \alpha)^{1/n}\}$$

where ' $\alpha$ ' is the level significance.

[Bhartiya Univ. M.Sc. (Maths.), 1991]

(b) Prove that the confidence interval for  $\theta$  obtained in (a) part above is shorter than the one obtained in Question 9 below.

8. Develop a general method for constructing confidence intervals. Consider a random sample of size  $n$  from the exponential distribution with p.d.f.

$$f(x, \theta) = e^{-(x-\theta)}, \theta \leq x < \infty, -\infty < \theta < \infty.$$

Show that  $P [X_{(1)} - \frac{1}{n} \log \alpha \leq \theta \leq X_{(1)}] = 1 - \alpha$

where symbols have their usual meanings. Also interpret the result.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

9. Consider a random sample  $X_1, X_2, \dots, X_n$  from an  $U[0, \theta]$  population. Show that  $R$  and  $R/\xi$  are the confidence limits for  $\theta$  with confidence coefficient  $(1 - \alpha)$ , where  $R$  is the sample range and  $\xi$  satisfies the equation :

$$\xi^{n-1} \{ n - (n-1)\xi \} = \alpha$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1985]

10. Explain the difference between point estimation and interval estimation.

Obtain 100  $(1 - \alpha)\%$  confidence interval for the population correlation coefficient ' $p$ ' when the random sample of size  $n$  has been drawn from bivariate normal population.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

11. Describe the pivotal quantity method for constructing confidence intervals.

Obtain a large sample 100  $(1 - \alpha)\%$  confidence interval for the parameter  $\theta$  in random sampling from the population :

$$dF(x) = \theta e^{-\theta x}; x > 0, \theta > 0$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

12. Develop a general method for obtaining confidence intervals. Obtain a 100  $(1 - \alpha)\%$  confidence interval for large sample size for the parameter  $\theta$  of the Poisson distribution :

$$f(x, \theta) = \frac{e^{-\theta} \theta^x}{x!}, x = 0, 1, 2, \dots$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

13. Describe the general method of constructing the confidence interval for large samples.

If  $X_1, X_2, \dots, X_n$  is a random sample from an exponential distribution with mean  $\theta$ , obtain 95% confidence interval for  $\theta$  when  $n$  is large.

[Delhi Univ. B.Sc. (Stat. Hons.), 1993]

14. (a) Show that with the exponential distribution

$$dF(x) = \theta e^{-\theta x}, x \geq 0$$

central confidence limits for  $\theta$  for large samples of size  $n$  and 95% confidence coefficient are :  $(1 \pm 1.96/\sqrt{n})/\bar{x}$ ,

where  $\bar{x}$  is the mean of the sample observations  $x_1, x_2, \dots, x_n$  drawn randomly from the exponential population.

[Indian Civil Services, 1983]

(b) Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with density function :  $f(x, \theta) = \theta e^{-\theta x}, 0 \leq x < \infty$

Find a  $100(1 - \alpha)$  (when  $0 < \alpha < 1$ ) percent confidence interval for the mean of this population, for large samples.

[*Madras Univ. B.Sc., 1991*]

15. (a) Discuss the problem of interval estimation. Obtain the minimum confidence interval for the variance for a random sample of size  $n$  from a normal population with unknown mean.

[*Indian Civil Services, 1991*]

(b) Give a method of determining the confidence limits for a single unknown parameter, stating the conditions of validity. From amongst intervals of Confidence Coefficient  $\alpha$ , how will you decide one as being superior to another ?

[*Indian Civil Services, 1989*]

16. Consider a random sample  $X_1, X_2, \dots, X_n$  from the exponential distribution with p.d.f.

$$f(x, \theta, p) = \frac{\exp(-x/\theta) \cdot x^{p-1}}{\Gamma_p \theta^p}, \quad x > 0$$

$$= 0 \quad \text{, otherwise}$$

If  $p$  is known, obtain a confidence interval for  $\theta$ , starting from the sufficient statistic  $\bar{X}/p$ .

**CHAPTER SIXTEEN**

## **Statistical Inference-II** *( Testing of Hypothesis, Non-parametric Methods and Sequential Analysis )*

---

**16.1. Introduction.** The main problems in statistical inference can be broadly classified into two areas :

- (i) The area of estimation of population parameters and setting up of confidence intervals for them, i.e., the area of *point and interval estimation* and
- (ii) *Tests of statistical hypothesis.*

The first topic has already been discussed in Chapter 15. In this chapter we shall discuss: (a) The theory of testing of hypothesis initiated by J. Neyman and E.S. Pearson (Section 16.2), (b) Sequential analysis propounded by A. Wald (Section 16.4) and (c) Non-parametric tests (Section 16.3). In Neyman-Pearson theory, we use statistical methods to arrive at decisions in certain situations where there is lack of certainty, on the basis of a sample whose size is fixed in advance while in Wald's sequential theory the sample size is not fixed but is regarded as a random variable. Before taking up a detailed discussion of the topics in (a), (b) and (c), we shall explain below certain concepts which are of fundamental importance.

**16.2. Statistical Hypothesis-Simple and Composite.** A *statistical hypothesis* is some statement or assertion about a population or equivalently about the probability distribution characterising a population which we want to verify on the basis of information available from a sample. If the statistical hypothesis specifies the population completely then it is termed as a *simple statistical hypothesis*, otherwise it is called a *composite statistical hypothesis*.

For example, if  $X_1, X_2, \dots, X_n$  is a random sample of size  $n$  from a normal population with mean  $\mu$  and variance  $\sigma^2$ , then the hypothesis

$$H_0: \mu = \mu_0, \sigma^2 = \sigma_0^2$$

is a simple hypothesis, whereas each of the following hypotheses is a composite hypothesis:

- |  |  |
|--|--|
| (i) $\mu = \mu_0$ , (ii) $\sigma^2 = \sigma_0^2$<br>(iii) $\mu < \mu_0, \sigma^2 = \sigma_0^2$<br>(v) $\mu = \mu_0, \sigma^2 < \sigma_0^2$ ,<br>(vii) $\mu < \mu_0, \sigma^2 > \sigma_0^2$ . | (iv) $\mu > \mu_0, \sigma^2 = \sigma_0^2$<br>(vi) $\mu = \mu_0, \sigma^2 > \sigma_0^2$ |
|--|--|

A hypothesis which does not specify completely ' $r$ ' parameters of a population is termed as a *composite hypothesis with r degrees of freedom*.

**16-2-1. Test of a Statistical Hypothesis.** A test of a statistical hypothesis is a two-action decision problem after the experimental sample values have been obtained, the two-actions being the acceptance or rejection of the hypothesis under consideration.

**16-2-2. Null Hypothesis.** In hypothesis testing, a statistician or decision-maker should not be motivated by prospects of profit or loss resulting from the acceptance or rejection of the hypothesis. He should be completely impartial and should have no brief for any party or company nor should he allow his personal views to influence the decision. Much, therefore, depends upon how the hypothesis is framed. For example, let us consider the 'light-bulbs' problem. Let us suppose that the bulbs manufactured under some standard manufacturing process have an average life of  $\mu$  hours and it is proposed to test a new procedure for manufacturing light bulbs. Thus, we have two populations of bulbs, those manufactured by standard process and those manufactured by the new process. In this problem the following three hypotheses may be set up :

- (i) New process is better than standard process.
- (ii) New process is inferior to standard process.
- (iii) There is no difference between the two processes.

The first two statements appear to be biased since they reflect a preferential attitude to one or the other of the two processes. Hence the best course is to adopt the hypothesis of no difference, as stated in (iii). This suggests that the statistician should take up the neutral or null attitude regarding the outcome of the test. His attitude should be on the null or zero line in which the experimental data has the due importance and complete say in the matter. This neutral or non-committal attitude of the statistician or decision-maker before the sample observations are taken is the keynote of the null hypothesis.

Thus, in the above example of light bulbs if  $\mu_0$  is the mean life (in hours) of the bulbs manufactured by the new process then the null hypothesis which is usually denoted by  $H_0$ , can be stated as follows :

$$H_0 : \mu = \mu_0.$$

As another example let us suppose that two different concerns manufacture drugs for inducing sleep, drug *A* manufactured by first concern and drug *B* manufactured by second concern. Each company claims that its drug is superior to that of the other and it is desired to test which is a superior drug *A* or *B*? To formulate the statistical hypothesis let *X* be a random variable which denotes the additional hours of sleep gained by an individual when drug *A* is given and let the random variable *Y* denote the additional hours of sleep gained when drug *B* is used. Let us suppose that *X* and *Y* follow the probability distributions with means  $\mu_X$  and  $\mu_Y$  respectively. Here our null hypothesis would be that there is no difference between the effects of two drugs. Symbolically,

$$H_0 : \mu_X = \mu_Y.$$

**16-2-3. Alternative Hypothesis.** It is desirable to state what is called an alternative hypothesis in respect of every statistical hypothesis being tested because the acceptance or rejection of null hypothesis is meaningful only when it is being tested against a rival hypothesis which should rather be explicitly mentioned. Alternative hypothesis is usually denoted by  $H_1$ . For

example, in the example of light bulbs, alternative hypothesis could be  $H_1 : \mu > \mu_0$  or  $\mu < \mu_0$  or  $\mu \neq \mu_0$ . In the example of drugs, the alternative hypothesis could be  $H_1 : \mu_x > \mu_y$  or  $\mu_x < \mu_y$  or  $\mu_x \neq \mu_y$ .

In both the cases, the first two of the alternative hypotheses give rise to what are called '*one tailed*' tests and the third alternative hypothesis results in '*two tailed*' tests.

**Important Remarks 1.** In the formulation of a testing problem and devising a 'test of hypothesis' the roles of  $H_0$  and  $H_1$  are not at all symmetric. In order to decide which one of the two hypotheses should be taken as null hypothesis  $H_0$  and which one as alternative hypothesis  $H_1$ , the intrinsic difference between the roles and the implications of these two terms should be clearly understood.

2. If a particular problem cannot be stated as a test between two simple hypotheses, i.e., simple null hypothesis against a simple alternative hypothesis, then the next best alternative is to formulate the problem as the test of a simple null hypothesis against a composite alternative hypothesis. In other words, one should try to structure the problem so that null hypothesis is simple rather than composite.

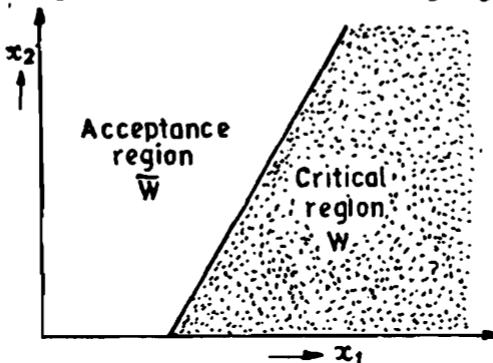
3. Keeping in mind the potential losses due to wrong decisions (which may or may not be measured in terms of money), the decision maker is somewhat conservative in holding the null hypothesis as true unless there is a strong evidence from the experimental sample observations that it is false. To him, the *consequences of wrongly rejecting a null hypothesis seem to be more severe than those of wrongly accepting it*. In most of the cases, the statistical hypothesis is in the form of a claim that a particular product or product process is superior to some existing standard. The null hypothesis  $H_0$  in this case is that there is no difference between the new product or production process and the existing standard. In other words, null hypothesis nullifies this claim. The rejection of the null hypothesis wrongly which amounts to the acceptance of claim wrongly involves huge amount of pocket expenses towards a substantive overhaul of the existing set-up. The resulting loss is comparatively regarded as more serious than the opportunity loss in wrongly accepting  $H_0$  which amounts to wrongly rejecting the claim, i.e., in sticking to the less efficient existing standard. In the light-bulbs problem discussed earlier, suppose the research division of the concern, on the basis of the limited experimentation, claims that its brand is more effective than that manufactured by standard process. If in fact, the brand fails to be more effective the loss incurred by the concern due to an immediate obsolescence of the product, decline of the concern's image, etc., will be quite serious. On the other hand, the failure to bring out a superior brand in the market is an opportunity loss and is not a consideration to be as serious as the other loss.

**16.2.4. Critical Region.** Let  $x_1, x_2, \dots, x_n$  be the sample observations denoted by  $O$ . All the values of  $O$  will be aggregate of a sample and they constitute a space, called the *sample space*, which is denoted by  $S$ .

Since the sample values  $x_1, x_2, \dots, x_n$  can be taken as a point in  $n$ -dimensional space, we specify some region of the  $n$ -dimensional space and see whether this point lies within this region or outside this region. We divide the

whole sample space  $S$  into two disjoint parts  $W$  and  $S - W$  or  $\bar{W}$  or  $W'$ . The null hypothesis  $H_0$  is rejected if the observed sample point falls in  $W$  and if it falls in  $W'$  we reject  $H_1$  and accept  $H_0$ . The region of rejection of  $H_0$  when  $H_0$  is true is that region of the outcome set where  $H_0$  is rejected if the sample point falls in that region and is called critical region. Evidently, the size of the critical region is  $\alpha$ , the probability of committing type 1 error (discussed below).

Suppose if the test is based on a sample of size 2, then the outcome set or the sample space is the first quadrant in a two-dimensional space and a test criterion will enable us to separate our outcome set into two complementary subsets,  $W$  and  $\bar{W}$ . If the sample point falls in the subset  $W$ ,  $H_0$  is rejected, otherwise  $H_0$  is accepted. This is shown in the following diagram :



**16-2-5. Two Types of Errors.** The decision to accept or reject the null hypothesis  $H_0$  is made on the basis of the information supplied by the observed sample observations. The conclusion drawn on the basis of a particular sample may not always be true in respect of the population. The four possible situations that arise in any test procedure are given in the following table.

#### DOUBLE DICHOTOMY RELATING TO DECISION AND HYPOTHESIS

		Decision From Sample	
		Reject $H_0$	Accept $H_0$
True State	$H_0$ True	Wrong (Type I Error)	Correct
		Correct	Wrong (Type II Error)

From the above table it is obvious that in any testing problem we are liable to commit two types of errors.

**Errors of Type I and Type II.** The error of rejecting  $H_0$  (accepting  $H_1$ ) when  $H_0$  is true is called *Type I error* and the error of accepting  $H_0$  when  $H_0$  is false ( $H_1$  is true) is called *Type II error*. The probabilities of type I and type II errors are denoted by  $\alpha$  and  $\beta$  respectively. Thus

- $\alpha = \text{Probability of type I error}$
- $= \text{Probability of rejecting } H_0 \text{ when } H_0 \text{ is true.}$
- $\beta = \text{Probability of type II error}$
- $= \text{Probability of accepting } H_0 \text{ when } H_0 \text{ is false.}$

Symbolically:

$$\left. \begin{aligned} P(x \in W | H_0) &= \alpha, \text{ where } x = (x_1, x_2, \dots, x_n) \\ \Rightarrow \int_W L_0 dx &= \alpha \end{aligned} \right\} \quad \dots(16.1)$$

where  $L_0$  is the likelihood function of the sample observations under  $H_0$  and  $\int dx$  represents the  $n$ -fold integral

$$\int \int \dots \int dx_1 dx_2 \dots dx_n.$$

Again

$$\left. \begin{aligned} P(x \in \bar{W} | H_1) &= \beta \\ \Rightarrow \int_{\bar{W}} L_1 dx &= \beta \end{aligned} \right\} \quad \dots(16.2)$$

where  $L_1$  is the likelihood function of the sample observations under  $H_1$ . Since

$$\int_W L_1 dx + \int_{\bar{W}} L_1 dx = 1,$$

we get

$$\int_W L_1 dx = 1 - \int_{\bar{W}} L_1 dx = 1 - \beta \quad \dots(16.2a)$$

$$\Rightarrow P(x \in W | H_1) = 1 - \beta \quad \dots(16.2b)$$

**16.2.6. Level of Significance.**  $\alpha$ , the probability of type I error, is known as the level of significance of the test. It is also called the *size of the critical region*.

**16.2.7. Power of the Test.**  $1 - \beta$ , defined in (16.2a) and (16.2b) is called the *power function of the test hypothesis  $H_0$  against the alternative hypothesis  $H_1$* . The value of the power function at a parameter point is called the *power of the test* at that point.

**Remarks 1.** In quality control terminology,  $\alpha$  and  $\beta$  are termed as *producer's risk* and *consumer's risk*, respectively.

**2.** An ideal test would be the one which properly keeps under control both the types of errors. But since the commission of an error of either type is a random variable, equivalently an ideal test should minimise the probability of both the types of errors, viz.,  $\alpha$  and  $\beta$ . But unfortunately, for a fixed sample size  $n$ ,  $\alpha$  and  $\beta$  are so related (like producer's and consumer's risk in sampling inspection plans), that the reduction in one results in an increase in the other. Consequently, the simultaneous minimising of both the errors is not possible.

Since type I error is deemed to be more serious than the type II error (c.f. Remark 3, § 16-2-3) the usual practice is to control  $\alpha$  at a predetermined low level and subject to this constraint on the probabilities of type I error, choose a test which minimises  $\beta$  or maximises the power function  $1 - \beta$ . Generally, we choose  $\alpha = 0.05$  or  $0.01$ .

**16-3. Steps in Solving Testing of Hypothesis Problem.** The major steps involved in the solution of a 'testing of hypothesis' problem may be outlined as follows :

1. Explicit knowledge of the nature of the population distribution and the parameter(s) of interest, i.e., the parameter(s) about which the hypotheses are set up.

2. Setting up of the null hypothesis  $H_0$  and the alternative hypothesis  $H_1$  in terms of the range of the parameter values each one embodies.

3. The choice of a suitable statistic  $t = t(x_1, x_2, \dots, x_n)$  called the *test statistic*, which will best reflect upon the probability of  $H_0$  and  $H_1$ .

4. Partitioning the set of possible values of the test statistic  $t$  into two disjoint sets  $W$  (called the *rejection region* or *critical region*) and  $\bar{W}$  (called the *acceptance region*) and framing the following test :

(i) Reject  $H_0$  (i.e., accept  $H_1$ ) if the value of  $t$  falls in  $W$ .

(ii) Accept  $H_0$  if the value of  $t$  falls in  $\bar{W}$ .

5. After framing the above test, obtain experimental sample observations, compute the appropriate test statistic and take action accordingly.

**16-4. Optimum Test Under Different Situations.** The discussion in § 16-3 and Remark 2, § 16-2-6 enables us to obtain the so called best test under different situations. In any testing problem the first two steps, viz., the form of the population distribution, the parameter(s) of interest and the framing of  $H_0$  and  $H_1$  should be obvious from the description of the problem. The most crucial step is the choice of the '*best test*', i.e., the best statistic ' $t$ ' and the critical region  $W$  where by *best test* we mean one which in addition to controlling  $\alpha$  at any desired low level has the minimum type II error  $\beta$  or maximum power  $1 - \beta$ , compared to  $\beta$  of all other tests having this  $\alpha$ . This leads to the following definition.

**16-4-1. Most Powerful Test (MP Test).** Let us consider the problem of testing a simple hypothesis

$$H_0: \theta = \theta_0$$

against a simple alternative hypothesis

$$H_1: \theta = \theta_1$$

**Definition.** The critical region  $W$  is the most powerful (MP) critical region of size  $\alpha$  (and the corresponding test a most powerful test of level  $\alpha$ ) for testing  $H_0: \theta = \theta_0$  against  $H_1: \theta = \theta_1$  if

$$P(x \in W | H_0) = \int_W L_0 dx = \alpha \quad \dots (16-3)$$

$$\text{and} \quad P(x \in W | H_1) \geq P(x \in W_1 | H_1) \quad \dots (16-3a)$$

for every other critical region  $W_1$  satisfying (16-3).

**16.4.2. Uniformly Most Powerful Test (UMP Test).** Let us now take up the case of testing a simple null hypothesis against a composite alternative hypothesis, e.g., of testing

$$H_0 : \theta = \theta_0$$

against the alternative

$$H_1 : \theta \neq \theta_0$$

In such a case, for a predetermined  $\alpha$ , the best test for  $H_0$  is called the *uniformly most powerful test of level  $\alpha$* .

**Definition.** The region  $W$  is called uniformly most powerful (UMP) critical region of size  $\alpha$  [and the corresponding test as uniformly most powerful (UMP) test of level  $\alpha$ ] for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta \neq \theta_0$  i.e.,  $H_1 : \theta = \theta_1 \neq \theta_0$  if

$$P(x \in W | H_0) = \int_W L_0 dx = \alpha \quad \dots(16.4)$$

$$\text{and} \quad P(x \in W | H_1) \geq P(x \in W_1 | H_1) \text{ for all } \theta \neq \theta_0, \quad \dots(16.4a)$$

whatever the region  $W_1$  satisfying (16.4) may be.

**16.5. Neyman J. and Pearson, E.S. Lemma.** This Lemma provides the most powerful test of simple hypothesis against a simple alternative hypothesis. The theorem, known as Neyman-Pearson Lemma, will be proved for density function  $f(x, \theta)$  of a single continuous variate and a single parameter. However, by regarding  $x$  and  $\theta$  as vectors, the proof can be easily generalised for any number of random variables  $x_1, x_2, \dots, x_n$  and any number of parameters  $\theta_1, \theta_2, \dots, \theta_k$ . The variables  $x_1, x_2, \dots, x_n$  occurring in this theorem are understood to represent a random sample of size  $n$  from the population whose density function is  $f(x, \theta)$ . The lemma is concerned with a simple hypothesis  $H_0 : \theta = \theta_0$  and a simple alternative  $H_1 : \theta = \theta_1$ .

**Theorem 16.1. (Neyman-Pearson Lemma).** Let  $k > 0$ , be a constant and  $W$  be a critical region of size  $\alpha$  such that

$$\begin{aligned} W &= \left\{ x \in S : \frac{f(x, \theta_1)}{f(x, \theta_0)} > k \right\} \\ \Rightarrow W &= \left\{ x \in S : \frac{L_1}{L_0} > k \right\} \quad \dots(16.5) \\ \text{and} \quad \overline{W} &= \left\{ x \in S : \frac{L_1}{L_0} \leq k \right\} \quad \dots(16.5a) \end{aligned}$$

where  $L_0$  and  $L_1$  are the likelihood functions of the sample observations  $x = (x_1, x_2, \dots, x_n)$  under  $H_0$  and  $H_1$  respectively. Then  $W$  is the most powerful critical region of the test hypothesis  $H_0 : \theta = \theta_0$  against the alternative  $H_1 : \theta = \theta_1$ .

**Proof.** We are given

$$P(x \in W | H_0) = \int_W L_0 dx = \alpha \quad \dots(16.6)$$

The power of the region is

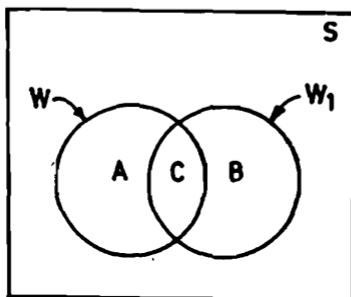
$$P(x \in W | H_1) = \int_W L_1 dx = 1 - \beta, \text{ (say).} \quad \dots (16-6a)$$

In order to establish the lemma, we have to prove that there exists no other critical region, of size less than or equal to  $\alpha$ , which is more powerful than  $W$ . Let  $W_1$  be another critical region of size  $\alpha_1 \leq \alpha$  and power  $1 - \beta_1$  so that we have

$$P(x \in W_1 | H_0) = \int_{W_1} L_0 dx = \alpha_1 \quad \dots (16-7)$$

$$\text{and} \quad P(x \in W_1 | H_1) = \int_{W_1} L_1 dx = 1 - \beta_1 \quad \dots (16-7a)$$

Now we have to prove that  $1 - \beta \geq 1 - \beta_1$



$$\text{Let } W = A \cup C \text{ and } W_1 = B \cup C$$

( $C$  may be empty, i.e.,  $W$  and  $W_1$  may be disjoint).

If  $\alpha_1 \leq \alpha$ , we have

$$\begin{aligned} & \int_{W_1} L_0 dx \leq \int_W L_0 dx \\ \Rightarrow & \int_{B \cup C} L_0 dx \leq \int_{A \cup C} L_0 dx \\ \Rightarrow & \int_B L_0 dx \leq \int_A L_0 dx \\ \Rightarrow & \int_A L_0 dx \geq \int_B L_0 dx \quad \dots (16-8) \end{aligned}$$

Since  $A \subset W$ ,

$$(16-5) \Rightarrow \int_A L_1 dx > K \int_A L_0 dx \geq k \int_B L_0 dx \quad \dots (16-8a)$$

[Using (16-8)]

Also [16.5 (a)] implies

$$\begin{aligned} \frac{L_1}{L_0} &\leq k \quad \forall x \in \bar{W} \\ \Rightarrow \int_{\bar{W}} L_1 dx &\leq k \int_{\bar{W}} L_0 dx \end{aligned}$$

This result also holds for any subset of  $\bar{W}$ , say  $\bar{W} \cap W_1 = B$ . Hence

$$\int_B L_1 dx \leq k \int_B L_0 dx \leq \int_A L_1 dx \quad [\text{From (16.8a)}]$$

Adding  $\int_C L_1 dx$  to both sides, we get

$$\begin{aligned} \int_{W_1} L_1 dx &\leq \int_W L_1 dx \\ \Rightarrow 1 - \beta &\geq 1 - \beta_1 \end{aligned}$$

Hence the Lemma.

**Remark.** Let  $W$  defined in (16.5) of the above theorem be the most powerful critical region of size  $\alpha$  for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , and let it be independent of  $\theta_1 \in \Theta_1 = \Theta - \Theta_0$ , where  $\Theta_0$  is the parameter space under  $H_0$ . Then we say that C.R.  $W$  is the UMP CR of size  $\alpha$  for testing  $H_0 : \theta = \theta_0$ , against  $H_1 : \theta \in \Theta_1$ .

**16.5.1. Unbiased Test and Unbiased Critical Region.** Let us consider the testing of  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ . The critical region  $W$  and consequently the test based on it is said to be unbiased if the power of the test exceeds the size of the critical region, i.e., if

$$\begin{aligned} \text{Power of the test} &\geq \text{size of the C.R.} & \dots (16.9) \\ \Rightarrow 1 - \beta &\geq \alpha \\ \Rightarrow P_{\theta_1}(W) &\geq P_{\theta_0}(W) \end{aligned}$$

$$\Rightarrow P[x : x \in W | H_1] \geq P[x : x \in W | H_0] \quad \dots (16.9a)$$

In other words, the critical region  $W$  is said to be unbiased if

$$P_{\theta}(W) \geq P_{\theta_0}(W), \forall \theta (\neq \theta_0) \in \Theta \quad \dots (16.9b)$$

**Theorem 16.2.** Every most powerful (MP) or uniformly most powerful (UMP) critical region (CR) is necessarily unbiased.

(i) If  $W$  be an MPCR of size  $\alpha$  for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , then it is necessarily unbiased.

(ii) Similarly if  $W$  be UMPCR of size  $\alpha$  for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta \in \Theta_1$ , then it is also unbiased.

**Proof.** Since  $W$  is an MPCR of size  $\alpha$  for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , by Neyman-Pearson Lemma, we have ; for  $\forall k > 0$ ,

$$W = \{x : L(x, \theta_1) \geq k L(x, \theta_0)\} = \{x : L_1 \geq k L_0\}$$

$$\text{and } W' = \{x : L(x, \theta_1) < k L(x, \theta_0)\} = \{x : L_1 < k L_0\},$$

where  $k$  is determined so that the size of the test is  $\alpha$  i.e.,

$$P_{\theta_0}(W) = P[x \in W | H_0] = \int_W L_0 dx = \alpha \quad \dots(i)$$

To prove that  $W$  is unbiased, we have to show that :

$$\text{Power of } W \geq \alpha \quad i.e., \quad P_{\theta_1}(W) \geq \alpha \quad \dots(ii)$$

We have :

$$P_{\theta_1}(W) = \int_W L_1 dx \geq k \int_{W'} L_0 dx = k\alpha$$

[ $\because$  On  $W$ ,  $L_1 \geq k L_0$  and Using (i)]

$$i.e., \quad P_{\theta_1}(W) \geq k\alpha, \quad \forall k > 0 \quad \dots(iii)$$

Also

$$1 - P_{\theta_1}(W) = 1 - P(x \in W | H_1) = P(x \in W' | H_1)$$

$$= \int_{W'} L_1 dx$$

$$< k \int_{W'} L_0 dx = k P(x : x \in W' | H_0)$$

[ $\because$  On  $W'$ ,  $L_1 < k L_0$ ]

$$= k [1 - P(x : x \in W | H_0)]$$

$$= k (1 - \alpha) \quad [\text{Using (i)}] \quad \dots(iv)$$

$$i.e., \quad 1 - P_{\theta_1}(W) < k (1 - \alpha), \quad \forall k > 0$$

Case (i)  $k \geq 1$ . If  $k \geq 1$ , then from (iii), we get

$$P_{\theta_1}(W) \geq k\alpha \geq \alpha$$

$\Rightarrow W$  is unbiased CR.

Case (ii)  $0 < k < 1$ . If  $0 < k < 1$ , then from (iv), we get :

$$1 - P_{\theta_1}(W) < 1 - \alpha$$

$$\Rightarrow P_{\theta_1}(W) > \alpha$$

$\Rightarrow W$  is unbiased C.R.

Hence MP critical region is unbiased.

(ii) If  $W$  is UMPCR of size  $\alpha$  then also the above proof holds if for  $\theta_1$  we write  $\theta$  such that  $\theta \in \Theta_1$ . So we have

$$P_{\theta}(W) > \alpha, \quad \forall \theta \in \Theta_1$$

$\Rightarrow W$  is unbiased CR.

**16-5-2. Optimum Regions and Sufficient Statistics.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a population with p.m.f. or p.d.f.  $f(x, \theta)$ , where the parameter  $\theta$  may be a vector. Let  $T$  be a sufficient statistic for  $\theta$ . Then by Factorization Theorem,

$$L(x, \theta) = \prod_{i=1}^n f(x_i, \theta) = g_{\theta}(t(x)) \cdot h(x) \quad \dots(*)$$

where  $g_{\theta}(t(x))$  is the marginal distribution of the statistic  $T = t(x)$ .

By Neyman-Pearson Lemma, the MPCR for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$  is given by :

$$W = \{x : L(x, \theta_1) \geq k L(x, \theta_0)\}, \forall k > 0 \quad \dots (**)$$

From (\*) and (\*\*), we get

$$\begin{aligned} W &= \{x : g_{\theta_1}(t(x)) \cdot h(x) \geq k \cdot g_{\theta_0}(t(x)) \cdot h(x)\}, \forall k > 0 \\ &= \{x : g_{\theta_1}(t(x)) \geq k \cdot g_{\theta_0}(t(x))\}, \forall k > 0 \end{aligned}$$

Hence if  $T = t(x)$  is sufficient statistic for  $\theta$  then the MPCR for the test may be defined in terms of the marginal distribution of  $T = t(x)$ , rather than the joint distribution of  $X_1, X_2, \dots, X_n$ .

**Example 16.1.** Given the frequency function :

$$\begin{aligned} f(x, \theta) &= \frac{1}{\theta}, \quad 0 \leq x \leq \theta \\ &= 0, \text{ elsewhere} \end{aligned}$$

and that you are testing the null hypothesis  $H_0 : \theta = 1$  against  $H_1 : \theta = 2$ , by means of a single observed value of  $x$ . What would be the sizes of the type I and type II errors, if you choose the interval (i)  $0.5 \leq x$ , (ii)  $1 \leq x \leq 1.5$  as the critical regions? Also obtain the power function of the test.

[Gauhati Univ. B.Sc. 1993; Calcutta Univ. B.Sc. (Maths Hons.), 1987]

**Solution.** Here we want to test

$$H_0 : \theta = 1, \text{ against } H_1 : \theta = 2.$$

$$(i) \text{ Here } W = \{x : 0.5 \leq x\} = \{x : x \geq 0.5\}$$

and

$$\begin{aligned} \bar{W} &= \{x : x \leq 0.5\} \\ \alpha &= P\{x \in \bar{W} | H_0\} = P\{x \geq 0.5 | \theta = 1\} \\ &= P(0.5 \leq x \leq 1 | \theta = 1) = P(0.5 \leq x \leq 1 | \theta = 1) \\ &= \int_{0.5}^1 [f(x, \theta)]_{\theta=1} dx = \int_{0.5}^1 1 \cdot dx = 0.5 \end{aligned}$$

Similarly,

$$\begin{aligned} \beta &= P\{x \in \bar{W} | H_1\} = P\{x \leq 0.5 | \theta = 2\} \\ &= \int_0^{0.5} [f(x, \theta)]_{\theta=2} dx = \int_0^{0.5} \frac{1}{2} dx = 0.25 \end{aligned}$$

Thus the sizes of type I and type II errors are respectively

$$\alpha = 0.5 \text{ and } \beta = 0.25$$

and power function of the test  $= 1 - \beta = 0.75$

$$(ii) \quad W = \{x : 1 \leq x \leq 1.5\}$$

$$\alpha = P\{x \in W | \theta = 1\} = \int_1^{1.5} [f(x, \theta)]_{\theta=1} dx = 0.$$

since under  $H_0 : \theta = 1, f(x, \theta) = 0$ , for  $1 \leq x \leq 1.5$ .

$$\beta = P\{x \in \bar{W} | \theta = 2\} = 1 - P\{x \in W | \theta = 2\}$$

$$= 1 - \int_1^{1.5} [f(x, \theta)]_{\theta=2} dx = 1 - \left| \frac{x}{2} \right|_1^{1.5} = 0.75$$

$\therefore$  Power Function =  $1 - \beta = 1 - 0.75 = 0.25$

**Example 16-2.** If  $x \geq 1$ , is the critical region for testing  $H_0 : \theta = 2$  against the alternative  $\theta = 1$ , on the basis of the single observation from the population,

$$f(x, \theta) = \theta \exp(-\theta x), 0 \leq x < \infty,$$

obtain the values of type I and type II errors.

[Poona Univ. M.C.A. 1993; Allahabad Univ. B.Sc., 1993;  
Delhi Univ. B.Sc (Stat. Hons.), 1988]

**Solution.** Here  $W = \{x : x \geq 1\}$  and  $\bar{W} = \{x : x < 1\}$ .

and  $H_0 : \theta = 2, H_1 : \theta = 1$

$\alpha$  = Size of Type I error

$$= P[x \in W | H_0] = P[x \geq 1 | \theta = 2]$$

$$= \int_1^{\infty} [f(x, \theta)]_{\theta=2} dx$$

$$= 2 \int_1^{\infty} e^{-2x} dx = 2 \left| \frac{e^{-2x}}{-2} \right|_1^{\infty}$$

$$= e^{-2} = 1/e^2$$

$\beta$  = Size of type II error

$$= P[x \in \bar{W} | H_1] = P[x < 1 | \theta = 1]$$

$$= \int_0^1 e^{-x} dx = \left| \frac{e^{-x}}{-1} \right|_0^1$$

$$= (1 - e^{-1}) = \frac{e - 1}{e}$$

**Example 16-3.** Let  $p$  be the probability that a coin will fall head in a single toss in order to test  $H_0 : p = \frac{1}{2}$  against  $H_1 : p = \frac{3}{4}$ . The coin is tossed 5 times and  $H_0$  is rejected if more than 3 heads are obtained. Find the probability of type I error and power of the test.

**Solution.** Here

$$H_0 : p = \frac{1}{2} \text{ and } H_1 : p = \frac{3}{4}.$$

If the r.v.  $X$  denotes the number of heads in  $n$  tosses of a coin then  $X \sim B(n, p)$  so that

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \binom{5}{x} p^x (1-p)^{5-x}, \quad \dots (*)$$

since  $n = 5$ , (given). The critical region is given by

$$W = \{x : x \geq 4\} \Rightarrow \bar{W} = \{x : x \leq 3\}$$

$\alpha_1$  = Probability of type I error

$$\doteq P[X \geq 4 | H_0]$$

$$= P[X = 4 | p = \frac{1}{2}] + P[X = 5 | p = \frac{1}{2}]$$

$$= \binom{5}{4} (\frac{1}{2})^4 (\frac{1}{2})^{5-4} + \binom{5}{5} (\frac{1}{2})^5 \quad [\text{From } (*)]$$

$$= 5(\frac{1}{2})^5 + (\frac{1}{2})^5 = 6(\frac{1}{2})^5$$

$$= \frac{3}{16}$$

$\beta$  = Probability of Type II error

$$= P[x \in \bar{W} | H_1] = 1 - P[x \in W | H_1]$$

$$= 1 - [P(X = 4 | p = \frac{3}{4}) + P(X = 5 | p = \frac{3}{4})]$$

$$= 1 - \left[ \binom{5}{4} (\frac{3}{4})^4 (\frac{1}{4}) + \binom{5}{5} (\frac{3}{4})^5 \right]$$

$$= 1 - (\frac{3}{4})^4 \left\{ \frac{5}{4} + \frac{3}{4} \right\}$$

$$= 1 - \frac{81}{128} = \frac{47}{128}$$

∴ Power of the test is

$$1 - \beta = \frac{81}{128}$$

**Example 16.4.** Let  $X \sim N(\mu, 4)$ ,  $\mu$  unknown. To test  $H_0 : \mu = -1$  against  $H_1 : \mu = 1$ , based on a sample of size 10 from this population, we use the critical region  $x_1 + 2x_2 + \dots + 10x_{10} \geq 0$ . What is its size? What is the power of the test?

**Solution.** Critical Region  $W = \{x : x_1 + 2x_2 + \dots + 10x_{10} \geq 0\}$ .

Let  $U = x_1 + 2x_2 + \dots + 10x_{10}$

Since  $x_i$ 's are i.i.d.  $N(\mu, 4)$ ,

$$U \sim N[(1+2+\dots+10)\mu, (1^2+2^2+\dots+10^2)\sigma^2] = N(55\mu, 385\sigma^2)$$

$$\Rightarrow U \sim N(55\mu, 385 \times 4) = N(55\mu, 1540) \quad \dots (*)$$

The size ' $\alpha$ ' of the critical region is given by :

$$\alpha = P(x \in W | H_0) = P(U \geq 0 | H_0) \quad \dots (**)$$

Under  $H_0 : \mu = -1$ ,  $U \sim N(-55, 1540)$

$$\Rightarrow Z = \frac{U - E(U)}{\sigma_U} = \frac{U + 55}{\sqrt{1540}}$$

$$\therefore \text{Under } H_0, \text{ when } U = 0, Z = \frac{55}{\sqrt{1540}} = \frac{55}{39.2428} = 1.4015$$

$$\begin{aligned}\therefore \alpha &= P(Z \geq 1.4015) && [\text{From } (**)] \\ &= 0.5 - P(0 \leq Z \leq 1.4015) \\ &= 0.5 - 0.4192 && (\text{From Normal Probability Tables}) \\ &= 0.0808\end{aligned}$$

Alternatively,  $\alpha = 1 - P(Z \leq 1.4015) = 1 - \Phi(1.4015)$ , where  $\Phi(\cdot)$  is the distribution function of standard normal variate.

Power of the test is given by :

$$1 - \beta = P(x \in W | H_1) = P(U \geq 0 | H_1)$$

Under  $H_1 : \mu = 1$ ,  $U \sim N(55, 1540)$

$$\Rightarrow Z = \frac{U - E(U)}{\sigma_U} = \frac{-55}{\sqrt{1540}} = -1.40 \quad (\text{when } U = 0)$$

$$\begin{aligned}\therefore 1 - \beta &= P(Z \geq -1.40) \\ &= P(-1.4 \leq Z \leq 0) + 0.5 \\ &= P(0 \leq Z \leq 1.4) + 0.5 \\ &= 0.4192 + 0.5 \\ &= 0.9192\end{aligned} \quad (\text{By symmetry})$$

Alternatively,

$$1 - \beta = 1 - P(Z \leq -1.40) = 1 - \Phi(-1.40).$$

**Example 16-5.** Let  $X$  have a p.d.f. of the form :

$$\begin{aligned}f(x, \theta) &= \frac{1}{\theta} e^{-x/\theta}; 0 < x < \infty, \theta > 0 \\ &= 0, \text{ elsewhere.}\end{aligned}$$

To test  $H_0 : \theta = 2$ , against  $H_1 : \theta = 1$ . use the random sample  $x_1, x_2$  of size 2, and define a critical region :

$$W = \{(x_1, x_2) : 9.5 \leq x_1 + x_2\}$$

Find : (i) Power of the test.

(ii) Significance level of the test.

**Solution.** We are given the critical region :

$$W = \{(x_1, x_2) : 9.5 \leq x_1 + x_2\} = \{(x_1, x_2) : x_1 + x_2 \geq 9.5\}$$

Size of the critical region i.e., the significance level of the test is given by :

$$\alpha = P(x \in W | H_0) = P[x_1 + x_2 \geq 9.5 | H_0] \quad ... (*)$$

In sampling from the given exponential distribution,

$$\frac{2}{\theta} \sum_{i=1}^n x_i \sim \chi^2_{(2n)} \quad [c.f. \text{ Example 16-8}]$$

$$\Rightarrow U = \frac{2}{\theta} (x_1 + x_2) \sim \chi^2_{(4)}, (n = 2).$$

$$\therefore \alpha = P \left[ \frac{2}{\theta} (x_1 + x_2) \geq \frac{2}{\theta} \times 9.5 \mid H_0 \right] \quad [\text{Form (*)}]$$

$$= P [\chi^2_{(4)} \geq 9.5] \quad (\because \text{Under } H_0, \theta = 2)$$

$\Rightarrow \alpha = 0.05$  [From Probability Tables of  $\chi^2$ -distribution]

Power of the test is given by

$$1 - \beta = P(x \in W \mid H_1) = P(x_1 + x_2 \geq 9.5 \mid H_1)$$

$$= P \left[ \frac{2}{\theta} (x_1 + x_2) \geq \frac{2}{\theta} \times 9.5 \mid H_1 \right]$$

$$= P [\chi^2_{(4)} \geq 19] \quad (\because \text{Under } H_1, \theta = 1)$$

**Example 16.6.** Use the Neyman-Pearson Lemma to obtain the best critical region for testing  $\theta = \theta_0$  against  $\theta = \theta_1 > \theta_0$  and  $\theta = \theta_1 < \theta_0$ , in the case of a normal population  $N(\theta, \sigma^2)$ , where  $\sigma^2$  is known. Hence find the power of the test.

[Delhi Univ. B.Sc. (Stat. Honors), 1986; Gujarat Univ. B.Sc. 1992]

**Solution.**

$$L = \prod_{i=1}^n f(x_i, \theta) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 \right]$$

Using Neyman-Pearson Lemma, the best critical region (B.C.R.) is given by (for  $k > 0$ )

$$\frac{L_1}{L_0} = \frac{\exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta_1)^2 \right]}{\exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta_0)^2 \right]} \geq k$$

$$\Rightarrow \exp \left[ -\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^n (x_i - \theta_1)^2 - \sum_{i=1}^n (x_i - \theta_0)^2 \right\} \right] \geq k$$

$$\Rightarrow \exp \left[ -\frac{n}{2\sigma^2} (\theta_1^2 - \theta_0^2) + \frac{1}{\sigma^2} (\theta_1 - \theta_0) \sum_{i=1}^n x_i \right] \geq k$$

$$\Rightarrow -\frac{n}{2\sigma^2} (\theta_1^2 - \theta_0^2) + \frac{1}{\sigma^2} (\theta_1 - \theta_0) \sum_{i=1}^n x_i \geq \log k$$

(since  $\log x$  is an increasing function of  $x$ )

$$\Rightarrow \bar{x}(\theta_1 - \theta_0) \geq \frac{\sigma^2}{n} \log k + \frac{\theta_1^2 - \theta_0^2}{2}$$

**Case (i)** If  $\theta_1 > \theta_0$ , the B.C.R. is determined by the relation (right-tailed test) :

$$\bar{x} > \frac{\sigma^2}{n} \cdot \frac{\log k}{\theta_1 - \theta_0} + \frac{\theta_1 + \theta_0}{2}$$

i.e.,  $\bar{x} > \lambda_1$ , (say).

$\therefore$  BCR is  $W = \{x : \bar{x} > \lambda_1\}$  ... (16.10)

Case (ii) If  $\theta_1 < \theta_0$ , the B.C.R. is given by the relation (left tailed test)

$$\bar{x} < \frac{\sigma^2}{n} \cdot \frac{\log k}{\theta_1 - \theta_0} + \frac{\theta_1 + \theta_0}{2} = \lambda_2, \text{ (say)}$$

Hence B.C.R. is  $W_1 = \{x : \bar{x} \leq \lambda_2\}$  ... (16.11)

The constants  $\lambda_1$  and  $\lambda_2$  are so chosen as to make the probability of each of the relations (16.10) and (16.11) equal to  $\alpha$  when the hypothesis  $H_0$  is true. The

sampling distribution of  $\bar{x}$ , when  $H_i$  is true is  $N\left(\theta_i, \frac{\sigma^2}{n}\right)$ , ( $i = 0, 1$ ).

Therefore the constants  $\lambda_1$  and  $\lambda_2$  are determined from the relations:

$$\begin{aligned} P[\bar{x} > \lambda_1 | H_0] &= \alpha, \text{ and } P[\bar{x} < \lambda_2 | H_0] = \alpha \\ \therefore P(\bar{x} > \lambda_1 | H_0) &= P\left[Z > \frac{\lambda_1 - \theta_0}{\sigma/\sqrt{n}}\right] = \alpha; Z \sim N(0, 1) \\ \Rightarrow \frac{\lambda_1 - \theta_0}{\sigma/\sqrt{n}} &= z_\alpha \Rightarrow \lambda_1 = \theta_0 + \frac{\sigma}{\sqrt{n}} z_\alpha \end{aligned} \quad \dots(16.12)$$

where  $z_\alpha$  is the upper  $\alpha$ -point of the standard normal variate given by

$$P(Z > z_\alpha) = \alpha \quad \dots(*)$$

Also  $P(\bar{x} < \lambda_2 | H_0) = \alpha \Rightarrow P(\bar{x} \geq \lambda_2 | H_0) = 1 - \alpha$

$$\begin{aligned} \Rightarrow P\left(Z \geq \frac{\lambda_2 - \theta_0}{\sigma/\sqrt{n}}\right) &= 1 - \alpha \Rightarrow \frac{\lambda_2 - \theta_0}{\sigma/\sqrt{n}} = z_{1-\alpha} \\ \Rightarrow \lambda_2 &= \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \end{aligned} \quad \dots(16.12a)$$

Note. By symmetry of normal distribution, we have  $z_{1-\alpha} = -z_\alpha$ .

**Power of the test.** By definition, the power of the test in case (i) is :

$$\begin{aligned} 1 - \beta &= P[x \in W | H_1] = P[\bar{x} \geq \lambda_1 | H_1] \\ &= P\left[Z \geq \frac{\lambda_1 - \theta_1}{\sigma/\sqrt{n}}\right] \quad [\because \text{Under } H_1, Z = \frac{\bar{x} - \theta_1}{\sigma/\sqrt{n}} \sim N(0, 1)] \\ &= P\left[Z \geq \frac{\theta_0 + \frac{\sigma}{\sqrt{n}} z_\alpha - \theta_1}{\sigma/\sqrt{n}}\right] \quad [\text{Using (16.12)}] \\ &= P\left[Z \geq z_\alpha - \frac{\theta_1 - \theta_0}{\sigma/\sqrt{n}}\right] \quad (\because \theta_1 > \theta_0) \end{aligned}$$

$$= 1 - P(Z \leq \lambda_3) \quad \left[ \lambda_3 = z_\alpha - \frac{\theta_1 - \theta_0}{\sigma/\sqrt{n}}, \text{ (say).} \right]$$

$$= 1 - \Phi(\lambda_3), \quad \dots(16-13)$$

where  $\Phi(\cdot)$  is the distribution function of standard normal variate.

Similarly in case (ii), ( $\theta_1 < \theta_0$ ), the power of the test is

$$1 - \beta = P(\bar{x} < \lambda_2 | H_1) = P\left(Z < \frac{\lambda_2 - \theta_1}{\sigma/\sqrt{n}}\right)$$

$$= P\left[Z < \frac{\theta_0 + \frac{\sigma}{\sqrt{n}}z_{1-\alpha} - \theta_1}{\sigma/\sqrt{n}}\right] \quad [\text{Using (16-12a)}]$$

$$= P\left[Z < z_{1-\alpha} + \frac{\theta_0 - \theta_1}{\sigma/\sqrt{n}}\right] \quad (\because \theta_0 > \theta_1)$$

$$= \Phi(\lambda_4), \quad \dots(16-13a)$$

$$\text{where } \lambda_4 = z_{1-\alpha} + \frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} = \frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} - z_\alpha \quad \dots(16-13b)$$

**UMP Critical Region.** (16-10) provides best critical region for testing  $H_0 : \theta = \theta_0$  against the hypothesis,  $H_1 : \theta = \theta_1$ , provided  $\theta_1 > \theta_0$  while (16-11) defines the best critical region for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , provided  $\theta_1 < \theta_0$ . Thus the best critical region for testing simple hypothesis  $H_0 : \theta = \theta_0$  against the simple hypothesis,  $H_1 : \theta = \theta_1 + c, c > 0$ , will not serve as the best critical region for testing simple hypothesis  $H_0 : \theta = \theta_0$  against simple alternative hypothesis  $H_1 : \theta = \theta_0 - c, c > 0$ .

Hence in this problem, no uniformly most powerful test exists for testing the simple hypothesis,  $H_0 : \theta = \theta_0$  against the composite alternative hypothesis,  $H_1 : \theta \neq \theta_0$ .

However, for each alternative hypothesis,  $H_1 : \theta = \theta_1 > \theta_0$  or  $H_1 : \theta = \theta_1 < \theta_0$ , a UMP test exists and is given by (16-10) and (16-11) respectively.

**Remark.** In particular, if we take  $n = 2$ , then the B.C.R. for testing  $H_0 : \theta = \theta_0$ , against  $H_1 : \theta = \theta_1 (> \theta_0)$  is given by : [From (16-10) and (16-12)]

$$W = \{x : (x_1 + x_2)/2 \geq \theta_0 + \sigma z_\alpha / \sqrt{2}\} \quad [\because \bar{x} = (x_1 + x_2)/2]$$

$$= \{x : x_1 + x_2 \geq 2\theta_0 + \sqrt{2}\sigma z_\alpha\}$$

$$= \{x : x_1 + x_2 \geq C\}, \text{ (say).} \quad \dots(**)$$

$$\text{where } C = 2\theta_0 + \sqrt{2}\sigma z_\alpha = 2\theta_0 + \sqrt{2}\sigma \times 1.645, \text{ if } \alpha = 0.05.$$

Similarly, the B.C.R. for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1 (< \theta_0)$  with  $n = 2$  and  $\alpha = 0.05$  is given by [From (16-11) and (16-12a)] :

$$\begin{aligned}
 W_1 &= \{x : (x_1 + x_2)/2 \leq \theta_0 - \sigma z_\alpha / \sqrt{2}\} \\
 &= \{x : (x_1 + x_2) \leq 2\theta_0 - \sqrt{2}\sigma \times 1.645\} \\
 &= \{x : x_1 + x_2 \leq C_1\}, \text{ (say)}, \quad \dots (***) \\
 \text{where } C_1 &= 2\theta_0 - \sqrt{2}\sigma z_\alpha = 2\theta_0 - \sqrt{2}\sigma \times 1.645.
 \end{aligned}$$

The B.C.R. for testing  $H_0 : \theta = \theta_0$  against the two tailed alternative  $H_1 : \theta = \theta_1 (\neq \theta_0)$ , is given by :

$$W_2 = \{x : (x_1 + x_2 \geq C) \cup (x_1 + x_2 \leq C_1)\} \quad \dots (****)$$

The regions in (\*\*), (\*\*\*) and (\*\*\*\*) are given by the shaded portions in the following figures (i), (ii) and (iii) respectively.

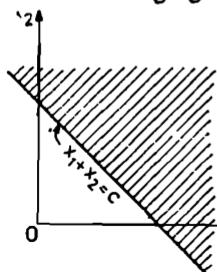


Fig. (i)

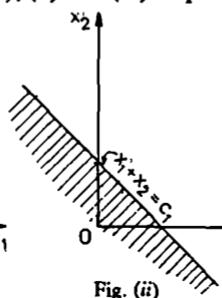


Fig. (ii)

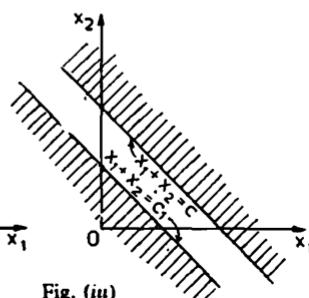


Fig. (iii)

$$\left. \begin{array}{l} \text{BCR for } H_0 : \theta = \theta_0 \\ \text{for } H_1 : \theta = \theta_1 (> \theta_0) \end{array} \right\} \quad \left. \begin{array}{l} \text{BCR for } H_0 : \theta = \theta_0 \\ \text{for } H_1 : \theta = \theta_1 (< \theta_0) \end{array} \right\} \quad \left. \begin{array}{l} \text{BCR for } H_0 : \theta = \theta_0 \\ \text{for } H_1 : \theta = \theta_1 (\neq \theta_0) \end{array} \right\}$$

**Example 16.7.** Show that for the normal distribution with zero mean and variance  $\sigma^2$ , the best critical region for  $H_0 : \sigma = \sigma_0$  against the alternative  $H_1 : \sigma = \sigma_1$  is of the form :

$$\sum_{i=1}^n x_i^2 \leq a_\alpha, \text{ for } \sigma_0 > \sigma_1$$

$$\text{and } \sum_{i=1}^n x_i^2 \geq b_\alpha, \text{ for } \sigma_0 < \sigma_1$$

Show that the power of the best critical region when  $\sigma_0 > \sigma_1$  is  $F\left(\frac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{\alpha, n}\right)$ , where  $\chi^2_{\alpha, n}$  is lower 100  $\alpha$ -per cent point and  $F(\cdot)$  is the distribution function of the  $\chi^2$ -distribution with  $n$  degrees of freedom.

**Solution.** Here we are given :

$$f(x, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right); -\infty < x < \infty, \sigma > 0.$$

The best critical region (B.C.R.), according to Neyman-Pearson Lemma, is given by (for  $k_\alpha > 0$ )

$$\frac{L_0}{L_1} \leq \frac{1}{k_\alpha} = A_\alpha, \text{ (say)}$$

$$\Rightarrow \left( \frac{\sigma_1}{\sigma_0} \right)^n \exp \left\{ - \frac{1}{2} \sum_{i=1}^n x_i^2 \left( \frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) \right\} \leq A_\alpha$$

$$\Rightarrow n \log \left( \frac{\sigma_1}{\sigma_0} \right) - \frac{1}{2} \sum_{i=1}^n x_i^2 \left( \frac{\sigma_1^2 - \sigma_0^2}{\sigma_0^2 \sigma_1^2} \right) \leq \log A_\alpha$$

(since  $\log x$  is an increasing function of  $x$ ).

$$\Rightarrow \frac{\sigma_0^2 - \sigma_1^2}{2\sigma_0^2 \sigma_1^2} \sum_{i=1}^n x_i^2 \leq \left[ \log A_\alpha - n \log \left( \frac{\sigma_1}{\sigma_0} \right) \right] \quad \dots (*)$$

**Case (i).** If  $\sigma_1 < \sigma_0$ , then B.C.R. is given by [From (\*)]

$$\sum_{i=1}^n x_i^2 \leq \left[ \log A_\alpha - n \log \left( \frac{\sigma_1}{\sigma_0} \right) \right] \frac{2\sigma_0^2 \sigma_1^2}{\sigma_0^2 - \sigma_1^2} = a_\alpha, \text{ (say).}$$

$$\text{i.e., } W = \left\{ \mathbf{x} : \sum_{i=1}^n x_i^2 \leq a_\alpha \right\}, \text{ for } \sigma_1 < \sigma_0. \quad \dots (16.14)$$

**Case (ii).** If  $\sigma_1 > \sigma_0$ , then B.C.R. is given by [From (\*)]

$$\sum_{i=1}^n x_i^2 \geq \left[ \log A_\alpha - n \log \left( \frac{\sigma_1}{\sigma_0} \right) \right] \cdot \frac{2\sigma_0^2 \sigma_1^2}{\sigma_0^2 - \sigma_1^2} = b_\alpha, \text{ (say).}$$

$$\text{i.e., } W_1 = \left\{ \mathbf{x} : \sum_{i=1}^n x_i^2 \geq b_\alpha \right\}, \text{ for } \sigma_1 > \sigma_0 \quad \dots (16.14a)$$

The constants  $a_\alpha$  and  $b_\alpha$  are so chosen that the size of the critical region is  $\alpha$ .Thus  $a_\alpha$  is determined so that  $P[\mathbf{x} \in W | H_0] = \alpha$ 

$$\Rightarrow P \left[ \sum_{i=1}^n x_i^2 \leq a_\alpha | H_0 \right] = \alpha$$

$$\Rightarrow P \left[ \sum_{i=1}^n \frac{x_i^2}{\sigma_0^2} \leq \frac{a_\alpha}{\sigma_0^2} | H_0 \right] = \alpha \quad \dots (**)$$

Since under  $H_0$ ,
$$\chi_{(n)}^2 = \sum_{i=1}^n \frac{x_i^2}{\sigma_0^2}$$
, is a  $\chi^2$ -variate with  $n$  d.f.,

$$\therefore P \left[ \chi_{(n)}^2 \leq \frac{a_\alpha}{\sigma_0^2} \right] = \alpha$$

$$\Rightarrow \frac{a_\alpha}{\sigma_0^2} = \chi^2_{\alpha, n} \Rightarrow \sigma_0^2 \chi^2_{\alpha, n} = a_\alpha \quad \dots (16.15)$$

where  $\chi^2_{\alpha, n}$  is the lower 100  $\alpha$ -per cent point of chi-square distribution with  $n$  d.f. given by

$$P(\chi^2 \leq \chi^2_{\alpha, n}) = \alpha \quad \dots (16.15a)$$

Hence the B.C.R. for testing  $H_0 : \sigma = \sigma_0$  against  $H_1 : \sigma = \sigma_1 (< \sigma_0)$ , is given by [From (16.14) and (16.15)] :

$$W = \left\{ \mathbf{x} : \sum_{i=1}^n x_i^2 \leq \sigma_0^2 \chi^2_{\alpha, n} \right\} \quad \dots(16.15b)$$

where  $\chi^2_{\alpha, n}$  is defined in (16.15a).

Also by definition, the power of the test is :

$$\begin{aligned} 1 - \beta &= P[\mathbf{x} \in W | H_1] = P\left[ \sum_{i=1}^n x_i^2 \leq a_\alpha | H_1 \right] \\ &= P\left[ \frac{\sum_{i=1}^n x_i^2}{\sigma_0^2} \leq \frac{a_\alpha}{\sigma_0^2} | H_1 \right] = P\left[ \frac{\sum_{i=1}^n x_i^2}{\sigma_0^2} \leq \chi^2_{\alpha, n} | H_1 \right] \\ &= P\left[ \frac{\sum_{i=1}^n x_i^2}{\sigma_1^2} \leq \frac{\sigma_0^2}{\sigma_1^2} \chi^2_{\alpha, n} | H_1 \right] \\ &= P\left[ \chi^2_{(n)} \leq \frac{\sigma_0^2}{\sigma_1^2} \chi^2_{\alpha, n} \right], \end{aligned}$$

since under  $H_1$ ,  $\sum x_i^2 / \sigma_1^2$  is a  $\chi^2$ -variate with  $n$  d.f.

$$\text{Hence, power of the test} = F\left(\frac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{\alpha, n}\right), \quad \dots(16.15c)$$

where  $F(\cdot)$  is the distribution function of chi-square distribution with  $n$  d.f.

**Remarks 1.** Similarly, for testing  $H_0 : \sigma = \sigma_0$  against  $H_1 : \sigma = \sigma_1 (> \sigma_0)$ ,  $b_\alpha$  in (16.14a) is determined so that :

$$\begin{aligned} P[\mathbf{x} \in W_1 | H_0] &= \alpha \\ \Rightarrow P\left[\mathbf{x} : \sum_{i=1}^n x_i^2 \geq b_\alpha | H_0\right] &= \alpha \\ \Rightarrow P\left[\mathbf{x} : \frac{\sum x_i^2}{\sigma_0^2} \geq \frac{b_\alpha}{\sigma_0^2} | H_0\right] &= \alpha \\ \Rightarrow P\left[\mathbf{x} : \chi^2_{(n)} \geq \frac{b_\alpha}{\sigma_0^2}\right] &= \alpha \\ \Rightarrow P\left[\mathbf{x} : \chi^2_{(n)} \leq \frac{b_\alpha}{\sigma_0^2}\right] &= 1 - \alpha \\ \Rightarrow \frac{b_\alpha}{\sigma_0^2} &= \chi^2_{1-\alpha, n} \Rightarrow b_\alpha = \sigma_0^2 \cdot \chi^2_{1-\alpha, n} \quad \dots(16.16) \end{aligned}$$

where  $\chi^2_{\alpha, n}$  is defined in (16.15a).

Hence the B.C.R. for testing  $H_0 : \sigma = \sigma_0$  against  $H_1 : \sigma = \sigma_1 (> \sigma_0)$ , is given by :

$$W_1 = \left\{ \mathbf{x} : \sum_{i=1}^n x_i^2 \geq \sigma_0^2 \cdot \chi^2_{1-\alpha, n} \right\} \quad \dots(16.16a)$$

The power of the test in this case is given by

$$\begin{aligned} 1 - \beta &= P(\mathbf{x} \in W_1 | H_1) = P \left[ \sum_{i=1}^n x_i^2 \geq \sigma_0^2 \cdot \chi^2_{1-\alpha, n} | H_1 \right] \\ &= P \left[ \frac{\sum_{i=1}^n x_i^2}{\sigma_1^2} \geq \frac{\sigma_0^2}{\sigma_1^2} \chi^2_{1-\alpha, n} | H_1 \right] \\ &= P \left[ \chi^2_{(n)} \geq \frac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{1-\alpha, n} \right], \end{aligned} \quad \dots(16.16b)$$

since under  $H_1$ ,  $\sum_{i=1}^n x_i^2 / \sigma_1^2$  is a  $\chi^2$ -variate with  $n$  d.f.

$$\begin{aligned} \therefore 1 - \beta &= 1 - P \left[ \chi^2_{(n)} \leq \frac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{1-\alpha, n} \right] \\ &= 1 - F \left( \frac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{1-\alpha, n} \right), \end{aligned} \quad \dots(16.16c)$$

where  $F(\cdot)$  is the distribution function of chi-square distribution with  $n$  d.f.

### 2. Graphical representation of the B.C.R. for the particular case $n = 2$ .

For  $n = 2$ , the B.C.R. for testing  $H_0 : \sigma = \sigma_0$ , against  $H_1 : \sigma = \sigma_1 (\sigma_1 < \sigma_0)$  is given by [From (16.15b)]

$$\begin{aligned} W &= \left\{ \mathbf{x} : \sum_{i=1}^2 x_i^2 \leq \sigma_0^2 \cdot \chi^2_{\alpha, 2} \right\} \\ &= \{ \mathbf{x} : x_1^2 + x_2^2 \leq a^2 \}, \end{aligned}$$

where  $a^2 = \sigma_0^2 \chi^2_{\alpha, 2}$ . Thus the B.C.R. is the interior of the circle with centre  $(0, 0)$  and radius ' $a$ ' and is shown as the shaded region in Figure (i) on page 16.22.

Similarly, from (16.16a), the B.C.R. for testing  $H_0 : \sigma = \sigma_0$ , against  $H_1 : \sigma = \sigma_1 (\sigma_1 > \sigma_0)$  for  $n = 2$  is given by :

$$W_1 = \{ \mathbf{x} : x_1^2 + x_2^2 \geq \sigma_0^2 \chi^2_{1-\alpha, 2} \} = \{ \mathbf{x} : x_1^2 + x_2^2 \geq b^2 \}$$

where  $b^2 = \sigma_0^2 \cdot \chi^2_{1-\alpha, 2}$ . Thus, B.C.R. is the exterior of the circle with centre  $(0, 0)$  and radius ' $b$ ' and is shown as the shaded region in Figure (ii) on page 16.22.

Similarly the B.C.R. for testing  $H_0 : \sigma = \sigma_0$  against the two-tailed alternative  $H_1 : \sigma = \sigma_1 (\neq \sigma_0)$ , for  $n = 2$  is given by :

$$\begin{aligned} W_3 &= W_1 \cup W_2 \\ &= \{ \mathbf{x} : x_1^2 + x_2^2 \leq a^2 \} \cup \{ \mathbf{x} : x_1^2 + x_2^2 \geq b^2 \} \end{aligned}$$

and is shown as the shaded region in the Figure (iii) below.

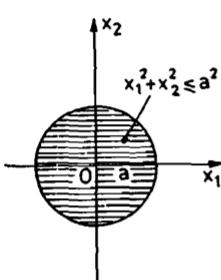


Fig. (i)

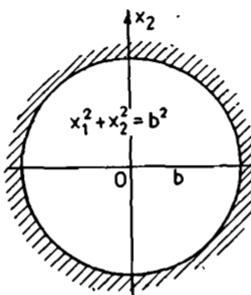


Fig. (ii)

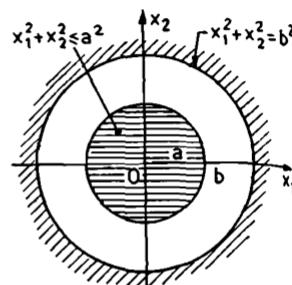


Fig. (iii)

3. (16-14) defines an UMP test for testing simple hypothesis  $H_0 : \sigma = \sigma_0$  against simple alternative hypothesis  $H_1 : \sigma = \sigma_1 (< \sigma_0)$  whereas (16-14a) defines an UMP test for testing simple hypothesis  $H_0 : \sigma = \sigma_0$  against the simple alternative hypothesis  $H_1 : \sigma = \sigma_1 (> \sigma_0)$ . However no UMP test exists for testing simple hypothesis  $H_0 : \sigma = \sigma_0$  against the composite alternative hypothesis  $H_1 : \sigma \neq \sigma_0$ .

**Example 16-8.** Given a random sample  $X_1, X_2, \dots, X_n$  from the distribution with p.d.f.  $f(x, \theta) = \theta e^{-\theta x}, x > 0$

show that there exists no UMP test for testing

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta \neq \theta_0.$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1988; Gorakhpur Univ. B.Sc., 1993]

Solution.  $L = \prod_{i=1}^n f(x_i, \theta) = \theta^n \cdot \exp \left[ -\theta \sum_{i=1}^n x_i \right]$

Consider

$$H_1 : \theta = \theta_1, (\theta_1 \neq \theta_0).$$

The best critical region, using Neyman-Pearson Lemma is given by :

$$\theta_1^n \exp [ -\theta_1 \sum x_i ] \geq k \cdot \theta_0^n \exp [ -\theta_0 \sum x_i ], k > 0.$$

$$\Rightarrow \exp [ (\theta_0 - \theta_1) \sum x_i ] \geq k \cdot \left( \frac{\theta_0}{\theta_1} \right)^n.$$

$$\Rightarrow (\theta_0 - \theta_1) \sum x_i \geq \log \left[ k \cdot \left( \frac{\theta_0}{\theta_1} \right)^n \right] = k_1, \text{ (say).}$$

... (\*)

Case (i) If  $\theta_1 > \theta_0$ , then B.C.R. is given by [From (\*)]

$$\sum x_i \leq \frac{k_1}{\theta_1 - \theta_0} = \lambda_1, \text{ (say).}$$

Case (ii) If  $\theta_1 < \theta_0$ , then B.C.R. is given by [From (\*)]

$$\sum x_i \geq \frac{k_1}{\theta_0 - \theta_1} = \lambda_2, \text{ (say).}$$

The constants  $\lambda_1$  and  $\lambda_2$  are so determined that

$$\begin{aligned} P[\sum x_i \leq \lambda_1 | H_0] &= \alpha & \text{and} & P[\sum x_i \geq \lambda_2 | H_0] &= \alpha \\ \Rightarrow P[2\theta \sum x_i \leq 2\theta \lambda_1 | H_0] &= \alpha & \Rightarrow P[2\theta \sum x_i \geq 2\theta \lambda_2 | H_0] &= \alpha \end{aligned}$$

But in random sampling from the given exponential distribution,

$$\begin{aligned} M_{\sum X_i}(t) &= \prod_{i=1}^n M_{X_i}(t) = [M_{X_i}(t)]^n \\ &= \left(1 - \frac{t}{\theta}\right)^n \\ \Rightarrow M_{2\theta \sum X_i}(t) &= M_{\sum X_i}(2t\theta) = (1 - 2t)^{-n}, \end{aligned}$$

which is the *m.g.f.* of a  $\chi^2$ -variate with  $2n$ . *d.f.* Hence by uniqueness theorem of *m.g.f.'s*,

$$2\theta \sum_{i=1}^n X_i \sim \chi^2_{(2n)}$$

Using this result in (\*\*)

$$\begin{aligned} P[2\theta_0 \sum X_i \leq \mu_1] &= P[\chi^2_{(2n)} \leq \mu_1] = \alpha \\ \Rightarrow \mu_1 &= \chi^2_{1-\alpha, 2n} \end{aligned}$$

where  $\chi^2_{\alpha, n}$  is the upper ' $\alpha$ ' point of  $\chi^2$ -distribution with *n.d.f.* given by

$$P(\chi^2 > \chi^2_{\alpha, n}) = \alpha \quad \dots(i)$$

Hence B.C.R. for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1 (> \theta_0)$  is given by

$$\begin{aligned} W_0 &= \{x : 2\theta_0 \sum X_i \leq \chi^2_{1-\alpha, 2n}\} \\ &= \{x : \sum X_i \leq \frac{1}{2\theta_0} \chi^2_{1-\alpha, 2n}\} \end{aligned}$$

and since it is independent of  $\theta_1$ ,  $W_0$  is *U.M.P.C.R.* for  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1 (> \theta_0)$ .

Similarly from (\*\*\*)<sup>1</sup>, we get

$$\begin{aligned} P[2\theta_0 \sum X_i \geq \mu_2] &= P[\chi^2_{(2n)} \geq \mu_2] = \alpha \\ \Rightarrow \mu_2 &= \chi^2_{\alpha, 2n} \end{aligned}$$

Hence B.C.R. for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1 (< \theta_0)$  is given by :

$$\begin{aligned} W_1 &= \{x : 2\theta_0 \sum X_i \geq \chi^2_{\alpha, 2n}\} \\ &= \{x : \sum X_i \geq \frac{1}{2\theta_0} \chi^2_{\alpha, 2n}\}, \end{aligned}$$

and since it is independent of  $\theta_1$ ,  $W_1$  is also *U.M.P.C.R.* for  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1 (< \theta_0)$ .

However, since the two critical regions  $W_0$  and  $W_1$  are different, there exists no critical region of size  $\alpha$  which is *U.M.P.* for  $H_0 : \theta = \theta_0$  against the two tailed alternative,  $H_1 : \theta \neq \theta_0$ .

*Power of the test.* The power of the test for testing  $H_0 : \theta = \theta_0$ , against  $H_1 : \theta = \theta_1 (> \theta_0)$  is given by

$$\begin{aligned} 1 - \beta &= P[x \in W_0 | H_1] \\ &= P\left[\sum_{i=1}^n X_i \leq \frac{1}{2\theta_0} \chi^2_{1-\alpha, 2n} | H_1\right] \end{aligned}$$

$$\begin{aligned}
 &= P \left[ 2\theta_1 \sum_{i=1}^n x_i \leq \frac{\theta_1}{\theta_0} \chi^2_{1-\alpha, 2n} \mid H_1 \right] \\
 &= P \left[ \chi^2_{(2n)} \leq \frac{\theta_1}{\theta_0} \chi^2_{1-\alpha, (2n)} \right], \quad \dots (*)
 \end{aligned}$$

since under  $H_1$ ,  $2\theta_1 \sum_{i=1}^n x_i \sim \chi^2_{(2n)}$ .

Similarly the power of the test for testing  $H_0 : \theta = \theta_0$ , against  $H_1 : \theta = \theta_1$  ( $< \theta_0$ ) is given by :

$$\begin{aligned}
 1 - \beta &= P [x \in W_1 \mid H_1] \\
 &= P \left[ \sum_{i=1}^n x_i \geq \frac{1}{2\theta_0} \chi^2_{\alpha, 2n} \mid H_1 \right] \\
 &= P \left[ 2\theta_1 \sum_{i=1}^n x_i \geq \frac{\theta_1}{\theta_0} \chi^2_{\alpha, 2n} \mid H_1 \right] \\
 &= P \left[ \chi^2_{(2n)} \geq \frac{\theta_1}{\theta_0} \chi^2_{\alpha, 2n} \right] \quad \dots (**)
 \end{aligned}$$

**Remark.** The graphic representation of the B.C.R. for  $H_0 : \theta = \theta_0$  against different alternatives  $H_1 : \theta = \theta_1 (> \theta_0)$ ,  $H_1 : \theta = \theta_1 (< \theta_0)$  and  $H_1 : \theta = \theta_1 (\neq \theta_0)$  for  $n = 2$ , can be done similarly as in Example 16.6, for the mean of normal distribution.

**Example 16.9.** For the distribution :

$$dF = \begin{cases} \beta \exp \{-\beta(x - \gamma)\} dx, & x \geq \gamma \\ 0, & x < \gamma \end{cases}$$

show that for a hypothesis  $H_0$  that  $\beta = \beta_0$ ,  $\gamma = \gamma_0$  and an alternative  $H_1$  that  $\beta = \beta_1$ ,  $\gamma = \gamma_1$ , the best critical region is the region given by

$$\bar{x} \leq \frac{1}{\beta_1 - \beta_0} \left\{ \gamma_1 \beta_1 - \gamma_0 \beta_0 - \frac{1}{n} \log k + \log \frac{\beta_1}{\beta_0} \right\}$$

provided that the admissible hypothesis is restricted by the condition

$$\gamma_1 \leq \gamma_0, \beta_1 \geq \beta_0 \quad (\text{Gauhati Univ. M.Sc., 1992})$$

$$\begin{aligned}
 \text{Solution. } f(x ; \beta, \gamma) &= \beta \exp \{-\beta(x - \gamma)\}, x \geq \gamma \\
 &= 0, \text{ otherwise}
 \end{aligned}$$

$$\begin{aligned}
 \therefore \prod_{i=1}^n f(x_i ; \beta, \gamma) &= \beta^n \exp \{-\beta \sum_{i=1}^n (x_i - \gamma)\}; x_1, x_2, \dots, x_n \geq \gamma \\
 &= 0, \text{ otherwise}
 \end{aligned}$$

Using Neyman-Pearson Lemma, the B.C.R. for  $k > 0$ , is given by

$$\begin{aligned}
 & \frac{\beta_1^n \exp \left[ -\beta_1 \sum_{i=1}^n (x_i - \gamma_1) \right]}{\beta_0^n \exp \left[ -\beta_0 \sum_{i=1}^n (x_i - \gamma_0) \right]} \geq k \\
 \Rightarrow & \left( \frac{\beta_1}{\beta_0} \right)^n \exp \left[ -\beta_1 \sum_{i=1}^n (x_i - \gamma_1) + \beta_0 \sum_{i=1}^n (x_i - \gamma_0) \right] \geq k \\
 \Rightarrow & \left( \frac{\beta_1}{\beta_0} \right)^n \exp [ -\beta_1 n(\bar{x} - \gamma_1) + \beta_0 n(\bar{x} - \gamma_0) ] \geq k \\
 \Rightarrow & n \log (\beta_1/\beta_0) - n\bar{x}(\beta_1 - \beta_0) + n\beta_1\gamma_1 - n\beta_0\gamma_0 \geq \log k \\
 & \quad (\text{since } \log x \text{ is an increasing function of } x) \\
 \Rightarrow & \bar{x}(\beta_1 - \beta_0) \leq \left\{ \gamma_1\beta_1 - \gamma_0\beta_0 - \frac{1}{n} \log k + \log \left( \frac{\beta_1}{\beta_0} \right) \right\} \\
 \Rightarrow & \bar{x} \leq \frac{1}{\beta_1 - \beta_0} \left\{ \gamma_1\beta_1 - \gamma_0\beta_0 - \frac{1}{n} \log k + \log \left( \frac{\beta_1}{\beta_0} \right) \right\}
 \end{aligned}$$

provided  $\beta_1 > \beta_0$ .

**Example 16-10.** Examine whether a best critical region exists for testing the null hypothesis  $H_0 : \theta = \theta_0$  against the alternative hypothesis  $H_1 : \theta = \theta_1 > \theta_0$  for the parameter  $\theta$  of the distribution :

$$f(x, \theta) = \frac{1 + \theta}{(x + \theta)^2}, \quad 1 \leq x < \infty$$

[Bangalore Univ. B.Sc., 1992]

$$\text{Solution. } \prod_{i=1}^n f(x_i, \theta) = (1 + \theta)^n \prod_{i=1}^n \frac{1}{(x_i + \theta)^2}$$

By Neyman-Pearson Lemma, the B.C.R. is given by

$$\begin{aligned}
 (1 + \theta_1)^n \prod_{i=1}^n \frac{1}{(x_i + \theta_1)^2} & \geq k (1 + \theta_0)^n \prod_{i=1}^n \frac{1}{(x_i + \theta_0)^2} \\
 \Rightarrow n \log (1 + \theta_1) - 2 \sum_{i=1}^n \log (x_i + \theta_1) & \geq \log k + n \log (1 + \theta_0) - 2 \sum_{i=1}^n \log (x_i + \theta_0) \\
 \Rightarrow 2 \sum_{i=1}^n \log \left( \frac{x_i + \theta_0}{x_i + \theta_1} \right) & \geq \log k + n \log \left( \frac{1 + \theta_0}{1 + \theta_1} \right)
 \end{aligned}$$

Thus the test criterion is  $\sum_{i=1}^n \log \left( \frac{x_i + \theta_0}{x_i + \theta_1} \right)$ , which cannot be put in the form of a function of the sample observations, not depending on the hypothesis. Hence no B.C.R. exists in this case.

### EXERCISE 16(a)

1. (a) What are simple and composite statistical hypotheses ? Give examples. Define null and alternative hypotheses. How is a statistical hypothesis tested ?

(b) Explain the following terms :

- (i) Errors of first and second kinds.
- (ii) The best critical region.
- (iii) Power function of a test.
- (iv) Level of significance.
- (v) Simple and composite hypotheses.
- (vi) Most powerful test.
- (vii) Uniformly most powerful test.

(c) Identify the composite hypotheses in the following, where  $\mu$  is the mean and  $\sigma^2$  is the variance of a distribution.

- (i)  $H_0 : \mu \leq 0, \sigma^2 = 1$
- (ii)  $H_0 : \mu = 0, \sigma^2 = 0$
- (iii)  $H_0 : \mu \leq 0, \sigma^2 = \text{arbitrary}$
- (iv)  $H_0 : \sigma^2 = \sigma_0^2$  (a given value),  $\mu$  arbitrary.

(d) (i) Explain the concepts of Type I and Type II errors, with examples and bring out their importance in Neyman and Pearson testing theory.

2. "In every hypothesis testing, the two types of errors are always present." If this is true then explain what is the use of hypothesis testing.

[Delhi Univ. M.C.A., 1990]

3. What is a statistical hypothesis ? Define (i) two types of errors, (ii) power of a test; with reference to testing of a hypothesis. Explain how the best critical region is determined. State clearly the theorem which is used to determine the best critical region for simple hypothesis at a given significance level.

[Calcutta Univ. B.Sc. (Maths. Hons.), 1992]

4. Explain the concept of the most powerful tests and discuss how the Neyman-Pearson lemma enables us to obtain the most powerful critical region for testing a simple hypothesis against a simple alternative.

[Madras Univ. B.Sc., 1988]

5. What is meant by a statistical hypothesis ? Explain the concepts of type I and type II errors. Show that a most powerful test is necessarily unbiased.

[Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1985]

6. What are simple and composite statistical hypotheses ? State and prove Neyman-Pearson Fundamental Lemma for testing a simple hypothesis against a simple alternative. [Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1986]

7. (a) Explain the basic concepts of statistical hypothesis. Discuss the problems associated with the testing of simple and composite hypotheses. Show that a most powerful test is necessarily unbiased.

[Delhi Univ. B.Sc. (Stat. Hons.), 1983]

8. State Neyman-Pearson Lemma.

Prove that if  $W$  is an MP region for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , then it is necessarily unbiased. Also prove that the same holds good if  $W$  is an UMP region. [Delhi Univ. B.Sc. (Stat. Hons.), 1982]

9. (a) Let  $p$  denote the probability of getting a head when a given coin is tossed once. Suppose that the hypothesis  $H_0 : p = 0.5$  is rejected in favour of  $H_1 : p = 0.6$  if 10 trials result in 7 or more heads. Calculate the probabilities of type I and type II errors. [Calcutta Univ. B.Sc. (Maths Hons.), 1989]

(b) An urn contains 6 marbles of which  $\theta$  are white and the others black. In order to test the null hypothesis  $H_0 : \theta = 3$ , against the alternative  $H_1 : \theta = 4$ , two marbles are drawn at random (without replacement) and  $H_0$  is rejected if both the marbles are white; otherwise  $H_0$  is accepted. Find the probabilities of committing type I and type II errors.

If it is decided to reject  $H_0$  when both marbles are black and to accept it otherwise, find the probabilities of rejecting  $H_0$  (i) when  $H_0$  is true and (ii) when  $H_1$  is true. Comment on your results.

10. (a)  $p$  is the probability that a given die shows even number. To test  $H_0 : p = \frac{1}{2}$  against  $H_1 : p = \frac{1}{3}$ , following procedure is adopted. Toss the die twice and accept  $H_0$  if both times it shows even number. Find the probabilities of type I and type II errors.

(b) Let  $p$  be the probability that a coin will fall head in a single toss. In order to test the hypothesis  $H_0 : p = \frac{1}{2}$ , the coin is tossed 6 times and the hypothesis  $H_0$  is rejected if more than 4 heads are obtained. Find the probability of the error of first kind. If the alternative hypothesis is  $H_1 : p = \frac{3}{4}$ , find the probability of the error of second kind.

(c) In a Bernoulli distribution with parameter  $p$ ,  $H_0 : p = \frac{1}{2}$  against  $H_1 : p = \frac{2}{3}$ , is rejected if more than 3 heads are obtained out of 5 throws of a coin. Find the probabilities of Type I and Type II errors.

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

11. (a) Let  $X_1, X_2, \dots, X_9$  be a random sample from  $N(\theta, 25)$ . If, for testing  $H_0 : \theta = 20$ , against  $H_1 : \theta = 26$ , the critical region  $W$  is defined by

$$W = \{x \mid \bar{X} > 23.266\},$$

then find the size of critical region and the power.

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

(b) Let  $X \sim N(\mu, 4)$ ,  $\mu$  unknown. To test  $H_0 : \mu = -1$  against  $H_1 : \mu = 1$ , based on a sample of size 10 from this population we use the critical region  $x_1 + 2x_2 + \dots + 10x_{10} \geq 0$ . What is its size? What is the power of the test?

(c) A sample of size 16 is drawn from a normal population with mean  $\mu$  and standard deviation  $\sigma$  for testing the hypothesis  $H_0 : \mu = \sigma = 1$ , against the alternative hypothesis  $H_1 : \mu = \sigma = 2$ . It is decided to reject the hypothesis  $H_0$  if the sample mean exceeds 1.5 and otherwise accept it.

Calculate the probabilities of errors of the first and second kind in this procedure.

$$\left[ \text{Given } \int_{-\infty}^1 \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 0.8413 \text{ and } \int_{-\infty}^2 \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 0.9773 \right]$$

12. (a) The hypothesis  $\mu = 50$ , is rejected if the mean of a sample of size 25 is either greater than 70.54 or less than 31.19. Assuming the distribution to be normal with s.d. 50, find the level of significance. Obtain the power function for the test and sketch the power curve with two values above 50 and two values below 50.

(b) Calculate the size of the type II error if the type I error is chosen to be  $\alpha = 0.16$  if you are testing  $H_0 : \mu = 7$  against  $H_1 : \mu = 6$ , for a normal distribution with  $\sigma = 2$ , by means of a sample of size 25 and if the proper tail of the  $\chi^2$  distribution is used as the critical region.

13. (a) Given the frequency function :

$$f(x, \theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta \\ = 0, \text{ elsewhere}$$

and that you are testing the hypothesis  $H_0 : \theta = 1.5$  against  $H_1 : \theta = 2.5$ , by means of a single observed value of  $x$ , what would be the sizes of the type I and type II errors, if you choose the interval  $0.8 \leq x$ , as the critical region? Also obtain the power function of the test.

(b) It is desired to test the hypothesis  $H_0 : \theta = 0$  against  $H_1 : \theta > 0$ , by observing a random variable  $X$  which is uniformly distributed on  $[\theta, \theta + 1]$ . Given only one observation, sketch the power function of the test whose critical region is defined by  $\{x > c\}$ . What value of  $c$  would you choose?

Given  $n$  observations, derive the general formula of the power function of the test whose critical region is defined by : (at least one  $x$  is greater than  $c$ ) and indicate how you would construct a confidence interval for  $\theta$ .

14. Let  $X$  have a p.d.f. of the form :

$$f(x, \theta) = \frac{1}{\theta} \exp(-x/\theta), \quad 0 < x < \infty, \theta > 0 \\ = 0, \text{ elsewhere.}$$

To test  $H_0 : \theta = 2$  against  $H_1 : \theta = 1$ , use a random sample  $X_1, X_2$  of size 2 and define a critical region  $C = \{(x_1, x_2) : 9.5 \leq x_1 + x_2\}$ .

Find (i) Power function of the test.

(ii) Significance level of the test .

(b) Let  $X$  have a p.d.f. of the form :

$$\begin{aligned} f(x; \theta) &= \theta x^{\theta-1}, 0 < x < 1 \\ &= 0, \text{ elsewhere} \end{aligned}$$

To test the simple hypothesis  $H_0 : \theta = 1$  against the alternative simple hypothesis  $H_1 : \theta = 2$ , use a random sample  $X_1, X_2$  of size  $n = 2$  and define the critical region to be

$$C = \{(x_1, x_2) : \frac{3}{4} \leq x_1 x_2\}$$

where  $x_1, x_2$  are the values assumed by a sample. Obtain the power function of the test.

[Madras Univ. B.Sc., Stat-Main, 1981]

Hint.  $Y = -\log X$  has an exponential distribution with parameter  $\theta$  i.e.,  $Y \sim \gamma(\theta, 1)$ .

15. (a) Give a working rule of finding the best critical region for testing a simple hypothesis against a simple alternative.

For a normal  $(m, \sigma^2)$  population with known  $\sigma$ , construct a test for the null hypothesis  $H_0 : m = m_0$  against the alternative  $m > m_0$ .

[Calcutta Univ. B.Sc., (Maths Hons.), 1989]

(b) Let  $(x_1, x_2, \dots, x_n)$  be a random sample from  $N(\theta, \sigma^2)$ , where  $\sigma^2$  is known. Obtain an UMP test for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta > \theta_0$ . Also find the power function of the test and examine if the test is unbiased.

[Delhi Univ. B.Sc. (Stat. Hons.), 1986, 1982]

16. (a) Obtain the most powerful test for testing the mean  $\mu = \mu_0$  against  $\mu = \mu_1$ , ( $\mu_1 > \mu_0$ ) when  $\sigma^2 = 1$  in normal population.

(b) Obtain the most powerful test of size  $\alpha$  for  $H_0 : \mu = \mu_0$  against  $H_1 : \mu = \mu_1$  when  $\mu_1 > \mu_0$ , if the probability density function of the random variable  $X$  is

$$f(x, \mu) = \frac{1}{\sqrt{8\pi}} \cdot \exp\left\{-\frac{1}{8}(x - \mu)^2\right\}, -\infty < x < \infty$$

17. (a) Let  $x_1, x_2, \dots, x_n$  denote a random sample from the distribution that has p.d.f.

$$f(x, \mu) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}(x - \mu)^2\right], -\infty < x < \infty$$

It is desired to test  $H_0 : \mu = 0$  against  $H_1 : \mu = 1$ .

(b) Let  $X_1, X_2, \dots, X_n$  denote a random sample from the normal distribution  $N(\theta, 1)$ ,  $\theta$  is unknown. Show that there is no uniformly most powerful test of the simple hypothesis  $H_0 : \theta = \theta_0$ , where  $\theta_0$  is a fixed number against the alternative composite hypothesis  $H_1 : \theta \neq \theta_0$ .

18. Let  $x_1, x_2, \dots, x_n$  be a random sample from  $N(\mu, \theta)$ , where  $\mu$  is known. Obtain an UMP test for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta < \theta_0$ . Also find the power function of the test. [Delhi Univ. B.Sc. (Stat. Hons.), 1985]

19. Define M.P. region and U.M.P. region. Show that an M.P. region is necessarily unbiased.

Obtain M.P. regions of size  $\alpha$  for testing—

$$(i) H_0 : \theta = \theta_0 \text{ against } H_1 : \theta = \theta_1, (\theta_1 > \theta_0)$$

$$(ii) H_0 : \theta = \theta_0 \text{ against } H_1 : \theta = \theta_1, (\theta_1 < \theta_0)$$

for  $N(\mu, \theta)$ , where  $\mu$  is known.

Show that the tests in (i) and (ii) are U.M.P. against one-sided alternative.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

20. (a) Let  $x_1, x_2, \dots, x_n$  be a random sample from a normal distribution  $N(0, \sigma^2)$ . Show that there exists a uniformly most powerful test with significance level  $\alpha$  for testing

$$H_0 : \sigma^2 = \sigma_1^2 \text{ against } H_1 : \sigma^2 < \sigma_1^2.$$

If  $n = 15$ ,  $\alpha = 0.05$  and  $\sigma_1^2 = 3$ , determine the best critical region and the power function of the above test. [Gujarat Univ. B.Sc., Oct. 1993]

(b) State Neyman and Pearson's fundamental lemma and apply it to obtain the test for testing  $\sigma^2 = 1$  against  $\sigma^2 > 1$ , when the sample is from  $N(0, \sigma^2)$ . Is this test UMP? Is it unbiased? Give reasons.

[Indian Civil Services (Main), 1990]

21. A sample of size 25 is drawn from a normal population with unknown mean  $\mu$  and variance 16. It is required to test the hypothesis  $H_0 : \mu = 1.0$  against the alternative  $H_1 : \mu = 3.0$  at 5% level of significance. Obtain the most powerful test for testing  $H_0$  against  $H_1$  and state how you will find its power. Is the test uniformly most powerful?

22. Explain the statistical procedure of testing the following hypothesis regarding the standard deviation ( $\sigma$ ) of normal population:

$$H_0 : \sigma = \sigma_0$$

$$H_1 : \sigma = \sigma_1 > \sigma_0$$

Will the test criterion remain the same when  $\sigma_1$  is changed to  $\sigma \neq \sigma_0$ ?

23. State and prove Neyman-Pearson Lemma. If  $x \geq 1$  is the critical region for testing  $H_0 : \theta = 2$  against the alternative  $H_1 : \theta = 1$ , on the basis of a single observation from the population

$$f(x, \theta) = \theta e^{-\theta x}, 0 \leq x < \infty, \theta > 0,$$

obtain the values of type I and type II errors and the power function of the test.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

(b) Given a random sample  $X_1, X_2, \dots, X_n$  of size  $n$  from the distribution with p.d.f.

$$f(x, \theta) = \theta e^{-\theta x}; x > 0, 0 < \theta < \infty,$$

show that UMP test for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta < \theta_0$  is given by

$$\left\{ x \mid \sum x_i \geq \frac{1}{2\theta_0} \chi_{a, 2n}^2 \right\}$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

(c) Explain the Neyman-Pearson Lemma for finding the best critical region for testing a simple hypothesis about the parameter  $\theta$  of the density function  $f(x, \theta)$ . Illustrate your answer by constructing the best critical region for

testing,  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1 < \theta_0$ , where  $\theta$  is the parameter of the distribution with p.d.f.,

$$f(x, \theta) = \theta e^{-\theta x}; 0 < x < \infty, \theta > 0.$$

[Meerut Univ. B.Sc., 1993; Poona Univ. B.Sc., Oct. 1991]

24. (a) Two independent observations  $x_1, x_2$  are made on a random variable  $X$  with density function :

$$f(x, \theta) = \frac{1}{\theta} \exp(-x/\theta); 0 < x < \infty, \theta > 0.$$

Test the null hypothesis  $H_0 : \theta = 2$  against the alternative  $H_1 : \theta = 4$ . If  $H_0$  is accepted when  $x_1 + x_2 < 9.5$ , and rejected otherwise, obtain the level of significance and power of the test.

(b) Let  $X_1$  be a random sample of size one from a population with p.d.f.  $f_\theta(x) = \frac{1}{\theta} e^{-x/\theta}$ ;  $x \geq 0, \theta > 0$ . Obtain : (i). the B.C.R. of size  $\alpha$  for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$  and (ii) the power of the test.

[Delhi Univ. B.Sc. (Stat. Hons.), 1983]

25. (a) Obtain the statistic for testing the hypothesis that the mean of a Poisson population is 2 against the alternative that it is 3, on the basis of  $n$  independent observations.

(b) Suppose you are testing  $H_0 : \lambda = 2$  against  $H_1 : \lambda = 1$ , where  $\lambda$  is the parameter of the Poisson distribution. Obtain the best critical region of the test.

26. (a) Suppose a random sample of size  $n$  is taken from the Poisson population  $\left( \frac{\exp(-\lambda) \cdot \lambda^x}{x!} \right)$ ,  $x = 0, 1, 2, \dots$ . Give the most powerful critical region of size  $\alpha$  for testing the hypothesis  $\lambda = \lambda_0$  against  $\lambda = \lambda_1$ , ( $\lambda_1 > \lambda_0$ ).

How can you use the above result to find a confidence interval for  $\lambda$  ?

Write an expression for the power function of the test for the hypothesis  $\lambda = \lambda_0$  against  $\lambda > \lambda_0$ .

(b)  $X_1, X_2, \dots, X_{10}$  is a random sample of size 10 from a Poisson distribution with mean  $\theta$ . Show that the critical region  $C$  defined by  $\sum_{i=1}^{10} x_i \geq 3$ , is the best critical region for testing  $H_0 : \theta = 0.1$  against  $H_1 : \theta = 0.5$ .

[Madras Univ. B.Sc., Oct. 1991]

27. (a) Let  $X_1, X_2, \dots, X_n$  denote a random sample from a distribution having p.d.f.

$$\begin{aligned} f(x, p) &= p^x (1-p)^{1-x}; x = 0, 1; 0 < p < 1 \\ &= 0, \text{ elsewhere} \end{aligned}$$

It is desired to test  $H_0 : p = \frac{1}{2}$  against  $H_1 : p = \frac{1}{3}$ .

(b) Suppose  $X$  has Bernoulli distribution with probability of success  $\theta$ . On the basis of a random sample of size  $n$  it is proposed to reject the null hypothesis,  $H_0 : \theta = \frac{1}{2}$  if

$$(X_1 + X_2 + \dots + X_n) \geq \frac{3}{8} \text{ or } \leq \frac{5}{8}$$

For  $n = 5$ , find the level of significance of the test.

28. (a) Let  $x_1, x_2, \dots, x_n$  be a random sample from a Bernoulli distribution with density :

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x}; x = 0, 1$$

Obtain a uniformly most powerful size- $\alpha$  test for  $H_0 : \theta = \theta_0$  against  $H_1 : \theta > \theta_0$ . Would you modify the test if  $H_1 : \theta < \theta_0$  ?

[Delhi Univ. M.A. (Eco.), 1987]

(b) The probability that a given machine produces a defective item is  $p$  and the quality of the items varies independently from one to another. Given a random sample of  $n = 20$  items produced by the machine, what is the form of the best acceptance region for testing  $H_0 : p = 0.05$  versus  $H_1 : p = 0.10$ ? What are the possible values of  $\alpha \leq 0.1$  (probability of type I error) in this case and the corresponding values of  $\beta$ , the probability of type II error?

29. Derive a most powerful test of the hypothesis  $\theta = \frac{1}{4}$  against the alternative  $\theta = \frac{1}{2}$  for the parameter  $\theta$  in a geometric distribution  $\theta (1 - \theta)^x$ ,  $x = 0, 1, 2, \dots$  based on a random sample of size 2.

30. Describe the method for finding the best critical region of size  $\alpha$  for testing a simple hypothesis against simple alternative one. Illustrate it by finding BCR for testing  $H_0 : \theta = 0$  against  $H_1 : \theta = 1$ , for the Cauchy distribution.

$$dF(x) = \frac{dx}{\pi[1 + (x - \theta)^2]}, -\infty < x < \infty$$

based on a random sample of size 1.

31. The distribution of  $x$  is :

$$\begin{aligned} f(x, \theta) &= \frac{1}{2}, \theta - 1 \leq x \leq \theta + 1 \\ &= 0, \text{ otherwise} \end{aligned}$$

If  $H_0 : \theta = 4$  and  $H_1 : \theta = 5$ , determine the critical region on the right hand tail of the distribution corresponding to  $\alpha = 0.25$ . Also calculate the probability of type II error.

[Kurukshetra Univ. M.A. (Eco.), 1992]

32. (a) Define simple and composite hypotheses. State and prove Neyman-Pearson Lemma.

(b) Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from p.d.f.

$$f(x, \theta) = \theta x^{\theta-1}, 0 < x < 1, \theta > 0.$$

Obtain the U.M.P. region of size  $\alpha$  for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta > \theta_0$ . Also find the power function of the test.

33. (a) Let  $x_1, x_2, \dots, x_n$  be  $n$  independent observations on a random variable  $X$  with density function

$$f(x, \theta) = \theta x^{\theta-1}; 0 < x < 1, \theta > 0$$

Show that the best critical region for testing  $H_0 : \theta = 1$  against  $H_1 : \theta = 2$ , can be defined in terms of the geometric mean of  $x_1, x_2, \dots, x_n$ .

(b) Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with p.d.f.

$$f(x, \theta) = \begin{cases} \theta x^{\theta-1}, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

where  $0 < \theta < \infty$ . Show that the M.P. test of level  $\alpha$  for testing  $H_0 : \theta = 1$  against the alternative  $H_1 : \theta = 2$ , is given by the critical region :

$$\left\{ x \mid \prod_{i=1}^n x_i > \exp \left[ -\frac{1}{2} \chi^2_{1-\alpha, 2n} \right] \right\}$$

where  $\chi^2_{1-\alpha, 2n}$  is the lower  $\alpha$ -point of the  $\chi^2$ -distribution with  $2n$  d.f.

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

34. Let  $X_1, X_2, \dots, X_n$  be a random sample from a p.d.f.

$$f(x, \theta) = \begin{cases} \theta x^{\theta-1}, & 0 \leq x \leq 1, \theta > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Find an U.M.P. test of size  $\alpha$  for testing  $H_0 : \theta = 1$  against  $H_1 : \theta > 1$ . Also obtain the power function. [Delhi Univ. B.Sc. (Stat. Hons.), 1992]

35. Let  $X_1, X_2, \dots, X_n$  be a random sample from discrete distribution with probability function  $f(x)$  for which  $x$  takes non-negative integral values  $0, 1, 2, \dots$ .

According to  $H_0$ :

$$\therefore f(x) = \begin{cases} \frac{e^{-1}}{x!}; & x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases}$$

According to  $H_1$ :

$$f(x) = \begin{cases} \frac{1}{2^{x+1}}; & x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases}$$

Obtain the critical region of the most powerful test of level  $\alpha$  for testing  $H_0$  against  $H_1$ . Also find the power of the test for the case  $n = 1$  and  $k = 1$ .

36.  $H_0$  denotes the null hypothesis that a given distribution has the p.d.f.

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, -\infty < x < \infty$$

and  $H_1$  denotes the alternative hypothesis that the distribution has the p.d.f.

$$\frac{1}{2} \exp(-|x|), -\infty < x < \infty.$$

Obtain the most powerful test for testing  $H_0$  against  $H_1$ .

37. It is required to test  $H_0$  against  $H_1$  from a single observation  $x$ , where  $H_0$  is the hypothesis that the p.d.f. is

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), (-\infty < x < \infty)$$

and  $H_1$  is the hypothesis that the p.d.f. is

$$f(x) = \frac{2}{\Gamma(1/4)} \exp(-x^4), (-\infty < x < \infty)$$

Obtain the most powerful test with level of significance  $\alpha$  in this case.

38. State Neyman-Pearson fundamental lemma. With the usual notations, if  $\beta$  is the power of the most powerful test of size  $\alpha$  for testing  $H_0 : p = p_0$  against  $H_1 : p = p_1$ , show that  $\alpha < \beta$  unless  $p_0 = p_1$ .

$$\text{If } p_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}, -\infty < x < \infty,$$

$$p_1(x) = \frac{1}{\pi} \cdot \frac{1}{1 + (x-\mu)^2}, -\infty < x < \infty$$

and  $\mu$  is known, determine the most powerful test of size  $\alpha$ . Calculate its power, if  $\alpha$  and  $\mu$  have specified values.

**16.6. Likelihood Ratio Test.** Neyman-Pearson Lemma based on the magnitude of the ratio of two probability density functions provides the best test for testing simple hypothesis against simple alternative hypothesis. The best test in any given situation depends on the nature of the population distribution and the form of the alternative hypothesis being considered. In this section we shall discuss a general method of test construction called the *Likelihood Ratio (L.R.) Test* introduced by Neyman and Pearson for testing a hypothesis, simple or composite, against a simple or composite alternative hypothesis. This test is related to the maximum likelihood estimates.

Before defining the test, we give below some notations and terminology.

**Parameter Space.** Let us consider a random variable  $X$  with p.d.f.  $f(x, \theta)$ . In most common applications, though not always, the functional form of the population distribution is assumed to be known except for the value of some unknown parameter(s)  $\theta$  which may take any value on a set  $\Theta$ . This is expressed by writing the p.d.f. in the form  $f(x, \theta), \theta \in \Theta$ . The set  $\Theta$ , which is the set of all possible values of  $\theta$  is called the *parameter space*. Such a situation gives rise not to one probability distribution but a family of probability distributions which we write as  $\{f(x, \theta), \theta \in \Theta\}$ . For example if  $X \sim N(\mu, \sigma^2)$ , then the parameter space

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$$

In particular, for  $\sigma^2 = 1$ , the family of probability distributions is given by

$$\{N(\mu, 1) : \mu \in \Theta\}, \text{ where } \Theta = \{\mu : -\infty < \mu < \infty\}$$

In the following discussion we shall consider a general family of distributions

$$\{f(x : \theta_1, \theta_2, \dots, \theta_k) : \theta_i \in \Theta, i = 1, 2, \dots, k\}$$

The null hypothesis  $H_0$  will state that the parameters belong to some subspace  $\Theta_0$  of the parameter space  $\Theta$ .

Let  $x_1, x_2, \dots, x_n$  be a random sample of size  $n > 1$  from a population with p.d.f.  $f(x, \theta_1, \theta_2, \dots, \theta_k)$ , where  $\Theta$ , the parameter space is the totality of all points that  $(\theta_1, \theta_2, \dots, \theta_k)$  can assume. We want to test the null hypothesis

$$H_0 : (\theta_1, \theta_2, \dots, \theta_k) \in \Theta_0$$

against all alternative hypotheses of the type

$$H_1 : (\theta_1, \theta_2, \dots, \theta_k) \in \Theta - \Theta_0$$

The likelihood function of the sample observations is given by

$$L = \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_k) \quad \dots(16-16)$$

According to the principle of maximum likelihood, the likelihood equation for estimating any parameter  $\theta_i$  is given by

$$\frac{\partial L}{\partial \theta_i} = 0, \quad (i = 1, 2, \dots, k) \quad \dots(16-17)$$

Using (16-17), we can obtain the maximum likelihood estimates for the parameters  $(\theta_1, \theta_2, \dots, \theta_k)$  as they are allowed to vary over the parameter space  $\Theta$  and the subspace  $\Theta_0$ . Substituting these estimates in (16-16), we obtain the maximum values of the likelihood function for variation of the parameters in  $\Theta$  and  $\Theta_0$  respectively. Then the criterion for the likelihood ratio test is defined as the quotient of these two maxima and is given by

$$\lambda = \lambda(x_1, x_2, \dots, x_n) = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \frac{\sup_{\theta \in \Theta_0} L(x, \theta)}{\sup_{\theta \in \Theta} L(x, \theta)}, \quad \dots(16-18)$$

where  $L(\hat{\Theta}_0)$  and  $L(\hat{\Theta})$  are the maxima of the likelihood function (16-16) with respect to the parameters in the regions  $\Theta_0$  and  $\Theta$  respectively.

The quantity  $\lambda$  is a function of the sample observations only and does not involve parameters. Thus  $\lambda$  being a function of the random variables, is also a random variable. Obvious  $\lambda > 0$ . Further

$$\Theta_0 \subset \Theta \Rightarrow L(\Theta_0) \leq L(\Theta) \Rightarrow \lambda \leq 1$$

Hence, we get

$$0 \leq \lambda \leq 1 \quad \dots(16-19)$$

The critical region for testing  $H_0$  (against  $H_1$ ) is an interval

$$0 < \lambda < \lambda_0, \quad \dots(16-20)$$

where  $\lambda_0$  is some number ( $< 1$ ) determined by the distribution of  $\lambda$  and the desired probability of type I error, i.e.,  $\lambda_0$  is given by the equation :

$$P(\lambda < \lambda_0 | H_0) = \alpha \quad \dots(16-21)$$

For example, if  $g(\cdot)$  is the p.d.f. of  $\lambda$  then  $\lambda_0$  is determined from the equation :

$$\int_0^{\lambda_0} g(\lambda | H_0) d\lambda = \alpha \quad \dots(16-21a)$$

A test that has critical region defined in (16-20) and (16-21) is a *likelihood ratio test* for testing  $H_0$ .

**Remark.** Equations (16-20) and (16-21) define the critical region for testing the hypothesis  $H_0$  by the likelihood ratio test. Suppose that the distribution of  $\lambda$  is not known but the distribution of some function of  $\lambda$  is known, then this knowledge can be utilized as given in the following theorem.

**Theorem 16-3.** If  $\lambda$  is the likelihood ratio for testing a simple hypothesis  $H_0$  and if  $U = \phi(\lambda)$  is a monotonic increasing (decreasing) function of  $\lambda$  then the test based on  $U$  is equivalent to the likelihood ratio test. The critical region for the test based on  $U$  is

$$\phi(0) < U < \phi(\lambda_0) \quad [\phi(\lambda_0) < U < \phi(0)] \quad \dots(16-22)$$

**Proof.** The critical region for the likelihood ratio test is given by  $0 < \lambda < \lambda_0$ , where  $\lambda_0$  is determined by

$$\int_0^{\lambda_0} g(\lambda | H_0) d\lambda = \alpha \quad \dots(*)$$

Let  $U = \phi(\lambda)$  be a monotonically increasing function of  $\lambda$ . Then (\*) gives

$$\alpha = \int_0^{\lambda_0} g(\lambda | H_0) d\lambda = \int_{\phi(0)}^{\phi(\lambda_0)} h(u | H_0) du$$

where  $h(u | H_0)$  is the p.d.f. of  $U$  when  $H_0$  is true. Here the critical region  $0 < \lambda < \lambda_0$  transforms to  $\phi(0) < U < \phi(\lambda_0)$ . However if  $U = \phi(\lambda)$  is a monotonic decreasing function of  $\lambda$ , then the inequalities are reversed and we get the critical region as  $\phi(\lambda_0) < U < \phi(0)$ .

2. If we are testing a simple null hypothesis  $H_0$  then there is a unique distribution determined for  $\lambda$ . But if  $H_0$  is composite, then the distribution of  $\lambda$  may or may not be unique. In such a case the distribution of  $\lambda$  may possibly be different for different parameter points in  $\Theta_0$  and then  $\lambda_0$  is to be chosen such that

$$\int_0^{\lambda_0} g(\lambda | H_0) d\lambda \leq \alpha \quad \dots(16-23)$$

for all values of the parameters in  $\Theta_0$ .

However, if we are dealing with large samples, a fairly satisfactory situation to this testing of hypothesis problem exists as stated (without proof) in the following theorem.

**Theorem 16-4.** Let  $x_1, x_2, \dots, x_n$  be a random sample from a population with p.d.f.  $f(x ; \theta_1, \theta_2, \dots, \theta_k)$  where the parameter space  $\Theta$  is  $k$ -dimensional. Suppose we want to test the composite hypothesis

$$H_0 : \theta_1 = \theta_1', \theta_2 = \theta_2', \dots, \theta_r = \theta_r' ; r < k$$

where  $\theta_1', \theta_2', \dots, \theta_r'$  are specified numbers. When  $H_0$  is true,  $-2 \log_e \lambda$  is asymptotically distributed as chi-square with  $r$  degrees of freedom, i.e., under

$$H_0, \quad -2 \log \lambda \sim \chi_{(r)}^2, \quad \text{if } n \text{ is large.} \quad \dots(16-24)$$

Since  $0 \leq \lambda \leq 1$ ,  $-2 \log \lambda$  is an increasing function of  $\lambda$  and approaches infinity when  $\lambda \rightarrow 0$ , the critical region for  $-2 \log \lambda$  being the right hand tail of the chi-square distribution. Thus at the level of significance ' $\alpha$ ', the test may be stated as follows :

$$\text{Reject } H_0 \text{ if } -2 \log \lambda > \chi_{(r)}^2(\alpha)$$

where  $\chi_{(r)}^2(\alpha)$  is the upper  $\alpha$ -point of the chi-square distribution with  $r$  d.f. given by

$$P[\chi^2 > \chi_{(r)}^2(\alpha)] = \alpha,$$

otherwise  $H_0$  may be accepted.

**16-6-1. Properties of Likelihood Ratio Test.** Likelihood ratio (*L.R.*) test principle is an intuitive one. If we are testing a simple hypothesis  $H_0$  against a simple alternative hypothesis  $H_1$ , then the *LR* principle leads to the same test as given by the Neyman-Pearson lemma. This suggests that *LR* test has some desirable properties, specially large sample properties.

In *LR* test, the probability of type I error is controlled by suitably choosing the cut off point  $\lambda_0$ . *LR* test is generally *UMP* if an *UMP* test at all exists. We state below, the two asymptotic properties of *LR* tests.

1. Under certain conditions,  $-2 \log \lambda$  has an asymptotic chi-square distribution.

2. Under certain assumptions, *LR* test is consistent.

**16-7.** In this section we shall illustrate how the likelihood ratio criterion can be used to obtain various standard tests of significance in Statistics.

**16-7-1. Test for the Mean of a Normal Population.** Let us take the problem of testing if the mean of a normal population has a specified value. Let  $(x_1, x_2, \dots, x_n)$  be a random sample of size  $n$  from the normal population with mean  $\mu$  and variance  $\sigma^2$ , where  $\mu$  and  $\sigma^2$  are unknown. Suppose we want to test the (composite) null hypothesis

$$H_0 : \mu = \mu_0 \text{ (specified), } 0 < \sigma^2 < \infty$$

against the composite alternative hypothesis

$$H_1 : \mu \neq \mu_0 ; 0 < \sigma^2 < \infty$$

In this case the parameter space  $\Theta$  is given by

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$$

and the subspace  $\Theta_0$  determined by the null hypothesis  $H_0$  is given by

$$\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, 0 < \sigma^2 < \infty\}$$

The likelihood function of the sample observations  $x_1, x_2, \dots, x_n$  is given by

$$L = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] \quad \dots(16.25)$$

The maximum likelihood estimates of  $\mu$  and  $\sigma^2$  are given by :

$$\left. \begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2 \end{aligned} \right\} \quad \dots(16.26)$$

Hence substituting in (16.25), the maximum of  $L$  in the parameter space  $\Theta$  is given by

$$L(\hat{\Theta}) = \left[ \frac{1}{2\pi s^2} \right]^{n/2} \cdot \exp \left( -\frac{n}{2} \right) \quad \dots(16.27)$$

In  $\Theta_0$ , the only variable parameter is  $\sigma^2$  and MLE of  $\sigma^2$  for given  $\mu = \mu_0$  is given by

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum (x_i - \mu_0)^2 = s_0^2, \text{ (say)} \\ &= \frac{1}{n} \sum (x_i - \bar{x} + \bar{x} - \mu_0)^2 \\ &= \frac{1}{n} \sum (x_i - \bar{x})^2 + (\bar{x} - \mu_0)^2, \end{aligned} \quad \dots(16.28)$$

the product term vanishes, since

$$\sum (x_i - \bar{x})(\bar{x} - \mu_0) = (\bar{x} - \mu_0) \sum (x_i - \bar{x}) = 0$$

$$\therefore \hat{\sigma}^2 = s^2 + (\bar{x} - \mu_0)^2 = s_0^2, \text{ (say).} \quad \dots(16.28a)$$

Hence substituting in (16.25), we get.

$$L(\hat{\Theta}_0) = \left[ \frac{1}{2\pi s_0^2} \right]^{n/2} \exp(-n/2) \quad \dots(16.28b)$$

The ratio of (16.28b) and (16.27) gives the likelihood ratio criterion

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left[ \frac{s^2}{s_0^2} \right]^{n/2} \quad \dots(16.29)$$

$$= \left[ \frac{s^2}{s^2 + (\bar{x} - \mu_0)^2} \right]^{n/2} = \left[ \frac{1}{1 + [(\bar{x} - \mu_0)^2 / s^2]} \right]^{n/2} \quad \dots(16.29a)$$

We have proved earlier (§ 14.2) that under  $H_0$ , the statistic

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

$$\text{where } S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{ns^2}{n-1},$$

follows Student's  $t$ -distribution with  $(n-1)$  d.f.

Thus

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} \sim t_{n-1} \quad \dots(16-30)$$

Substituting in (16-29a), we get

$$\lambda = \frac{1}{\left(1 + \frac{t^2}{n-1}\right)^{n/2}} = \phi(t^2), \text{ (say).} \quad \dots(16-31)$$

The likelihood ratio test for testing  $H_0$  against  $H_1$  consists in finding a critical region of the type  $0 < \lambda < \lambda_0$ , where  $\lambda_0$  is given by (16-21a), which requires the distribution of  $\lambda$  under  $H_0$ . In this case, it is not necessary to obtain the distribution of  $\lambda$  since  $\lambda = \phi(t^2)$  is a monotonic function of  $t^2$  and the test can well be carried on with  $t^2$  as a criterion as with  $\lambda$  [c.f. Theorem 16-1]. Now  $t^2 = 0$  when  $\lambda = 1$  and  $t^2$  becomes infinite when  $\lambda = 0$ . The critical region of the LR test viz.,  $0 < \lambda < \lambda_0$ , on using (16-31) is equivalent to

$$\begin{aligned} & \left(1 + \frac{t^2}{n-1}\right)^{-n/2} \leq \lambda_0 \\ \Rightarrow & \left(1 + \frac{t^2}{n-1}\right)^{n/2} \geq \lambda_0^{-1} \\ \Rightarrow & \frac{t^2}{n-1} \geq (\lambda_0)^{-2/n} - 1 \\ \Rightarrow & t^2 \geq (n-1) [\lambda_0^{-2/n} - 1] = A^2, \text{ (say).} \end{aligned}$$

Thus the critical region may well be defined by

$$|t| = \left| \frac{\sqrt{n}(\bar{x} - \mu_0)}{S} \right| \geq A \quad \dots(16-32)$$

where the constant  $A$  is determined such that

$$P[|t| \geq A | H_0] = \alpha \quad \dots(16-33)$$

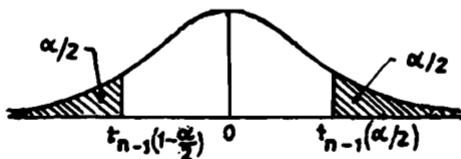
Since under  $H_0$ , the statistic  $t$  follows Student's  $t$ -distribution with  $(n-1)$  d.f.,

$$A = t_{n-1}(\alpha/2)$$

where the symbol  $t_n(\alpha)$  stands for the *right tail* 100  $\alpha\%$  point of the  $t$ -distribution with  $n$  d.f. given by

$$P\{t > t_n(\alpha)\} = \int_{t_n(\alpha)}^{\infty} f(t) dt = \alpha \quad \dots(16-33a)$$

where  $f(\cdot)$  is the p.d.f. of Student's  $t$  with  $n$  d.f. The critical region is shown in the following diagram.



Thus for testing  $H_0 : \mu = \mu_0$  against  $\mu \neq \mu_0$  ( $\sigma^2$ -unknown), we have the two-tailed  $t$ -test defined as follows :

If  $|t| = \left| \frac{\sqrt{n}(\bar{x} - \mu_0)}{S} \right| > t_{n-1}(\alpha/2)$ , reject  $H_0$  and if  $|t| < t_{n-1}(\alpha/2)$ ,  $H_0$  may be accepted.

**Important Remarks.** 1. Let us now consider the problem of testing the hypothesis

$$H_0 : \mu = \mu_0, 0 < \sigma^2 < \infty$$

against the alternative hypothesis

$$H_1 : \mu > \mu_0, 0 < \sigma^2 < \infty$$

Here  $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$

and  $\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, 0 < \sigma^2 < \infty\}$

The maximum likelihood estimates of  $\mu$  and  $\sigma^2$  belonging to  $\Theta$  are given by

$$\hat{\mu} = \begin{cases} \bar{x}, & \text{if } \bar{x} \geq \mu_0 \\ \mu_0, & \text{if } \bar{x} < \mu_0. \end{cases} \quad \dots(16-34)$$

$$\text{and } \hat{\sigma}^2 = \begin{cases} s^2, & \text{if } \bar{x} \geq \mu_0 \\ s_0^2, & \text{if } \bar{x} < \mu_0 \end{cases} \quad \dots(16-34a)$$

$$\text{where } s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2 \quad \dots(16-34b)$$

Thus

$$L(\hat{\Theta}) = \begin{cases} \left(\frac{1}{2\pi s^2}\right)^{n/2} \exp\left(-\frac{n}{2}\right), & \text{if } \bar{x} \geq \mu_0 \\ \left(\frac{1}{2\pi s_0^2}\right)^{n/2} \cdot \exp\left(-\frac{n}{2}\right), & \text{if } \bar{x} < \mu_0 \end{cases} \quad \dots(16.35)$$

In  $\Theta_0$ , the only unknown parameter is  $\sigma^2$  whose MLE is given by  $\hat{\sigma}^2 = s_0^2$ . Thus

$$L(\hat{\Theta}_0) = \left(\frac{1}{2\pi s_0^2}\right)^{n/2} \exp\left(-\frac{n}{2}\right) \quad \dots(16.36)$$

$$\therefore \lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \begin{cases} (s^2/s_0^2)^{n/2}, & \text{if } \bar{x} \geq \mu_0 \\ 1 & \text{if } \bar{x} < \mu_0 \end{cases} \quad \dots(16.37)$$

Thus the sample observations  $(x_1, x_2, \dots, x_n)$  for which  $\bar{x} < \mu_0$  are to be included in the acceptance region. Hence for the sample observations for which  $\bar{x} \geq \mu_0$ , the likelihood ratio criterion becomes

$$\lambda = (s^2/s_0^2)^{n/2}, \bar{x} \geq \mu_0 \quad \dots(16.37a)$$

which is the same as the expression obtained in (16.29). Proceeding similarly as in the above problem, the critical region of the form  $0 < \lambda < \lambda_0$  will be equivalently given by [c.f. (16.32)]

$$t^2 = \frac{n(\bar{x} - \mu_0)^2}{s^2} \geq A^2$$

$$\text{or by} \quad t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \geq A \quad (\because \bar{x} \geq \mu_0) \quad \dots(16.38)$$

where  $t$  follows Student's t distribution with  $(n - 1)$  d.f. The constant  $A$  is to be determined so that

$$P(t > A) = \alpha \quad \dots(16.39)$$

$$\Rightarrow A = t_{n-1}(\alpha)$$

Hence for testing  $H_0 : \mu = \mu_0$  against  $H_1 : \mu > \mu_0$ , we have the right-tailed-t-test defined as follows :

$$\text{Reject } H_0 \text{ if } t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} > t_{n-1}(\alpha) \text{ and}$$

$$\text{if } t < t_{n-1}(\alpha), H_0 \text{ may be accepted.}$$

2. If we want to test

$$H_0 : \mu = \mu_0, 0 < \sigma^2 < \infty$$

against the alternative hypothesis

$$H_1 : \mu < \mu_0, 0 < \sigma^2 < \infty,$$

then proceeding exactly similarly as in Remark 1 above, we shall get the critical region given by

$$t < -t_{n-1}(\alpha) \quad \dots(16-40)$$

In this case we have the left tailed t-test defined as follows :

If  $t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{S} < -t_{n-1}(\alpha)$ , reject  $H_0$  otherwise  $H_0$  may be accepted.

3. We summarise below in a tabular form the test criterion, along with the confidence interval for the parameter for testing the hypothesis  $H_0 : \mu = \mu_0$  against various alternatives for the normal population when  $\sigma^2$  is not known.

[Here  $t_n(\alpha)$  is upper  $\alpha$ -point of the t-distribution with  $n$  d.f. as defined in (16-33a).]

### NORMAL POPULATION $N(\mu, \sigma^2)$ ; $\sigma^2$ UNKNOWN

Serial No.	Hypothesis	Test	Test Statistic	Reject $H_0$ at Level of Significance, $\alpha$ , if	$(1 - \alpha)$ confidence interval for $\mu$
1.	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	Two tailed test	$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$	$ t  > t_{n-1}(\alpha/2)$	$\bar{x} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2) \leq \mu$ $\leq \bar{x} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)$
2.	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	Right tailed test	— do —	$t > t_{n-1}(\alpha)$	$\mu \geq \bar{x} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha)$
3.	$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	Left tailed test	— do —	$t < -t_{n-1}(\alpha)$	$\mu \leq \bar{x} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha)$

**16-7-2. Test for the Equality of Means of Two Normal Populations.** Let us consider two independent random variables  $X_1$  and  $X_2$  following normal distributions  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$  respectively where the means  $\mu_1, \mu_2$  and the variances  $\sigma_1^2, \sigma_2^2$  are unspecified. Suppose we want to test the hypothesis :

$H_0 : \mu_1 = \mu_2 = \mu$ , (say), (unspecified);  $0 < \sigma_1^2 < \infty, 0 < \sigma_2^2 < \infty$ ,  
against the alternative hypothesis

$$H_1 : \mu_1 \neq \mu_2, \sigma_1^2 > 0, \sigma_2^2 > 0.$$

**Case I. Population variances are unequal.**

$$\Theta = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_i < \infty, \sigma_i^2 > 0, i = 1, 2\}$$

and

$$\Theta_0 = \{(\mu, \sigma_1^2, \sigma_2^2) : -\infty < \mu < \infty, \sigma_i^2 > 0, i = 1, 2\}$$

Let  $x_{1i}$  ( $i = 1, 2, \dots, m$ ) and  $x_{2j}$  ( $j = 1, 2, \dots, n$ ) be two independent random samples of sizes  $m$  and  $n$  from the populations  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$  respectively. Then the likelihood function is given by

$$L = \left( \frac{1}{2\pi\sigma_1^2} \right)^{m/2} \cdot \exp \left[ -\frac{1}{2\sigma_1^2} \sum_{i=1}^m (x_{1i} - \mu_1)^2 \right] \\ \times \left( \frac{1}{2\pi\sigma_2^2} \right)^{n/2} \cdot \exp \left[ -\frac{1}{2\sigma_2^2} \sum_{j=1}^n (x_{2j} - \mu_2)^2 \right] \quad \dots(16-41)$$

The maximum likelihood estimates for  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1^2$  and  $\sigma_2^2$  are given by the equations :

$$\left. \begin{aligned} \frac{\partial}{\partial \mu_1} \log L &= 0 \Rightarrow \hat{\mu}_1 = \frac{1}{m} \sum_{i=1}^m x_{1i} = \bar{x}_1, \\ \frac{\partial}{\partial \mu_2} \log L &= 0 \Rightarrow \hat{\mu}_2 = \frac{1}{n} \sum_{j=1}^n x_{2j} = \bar{x}_2 \\ \frac{\partial}{\partial \sigma_1^2} \log L &= 0 \Rightarrow \hat{\sigma}_1^2 = \frac{1}{m} \sum_{i=1}^m (x_{1i} - \bar{x}_1)^2 = s_1^2, \text{ (say).} \\ \text{and } \frac{\partial}{\partial \sigma_2^2} \log L &= 0 \Rightarrow \hat{\sigma}_2^2 = \frac{1}{n} \sum_{j=1}^n (x_{2j} - \bar{x}_2)^2 = s_2^2, \text{ (say).} \end{aligned} \right\} \dots(16-41a)$$

Substituting in (16-41), we get

$$L(\Theta) = \left( \frac{1}{2\pi s_1^2} \right)^{m/2} \cdot \left( \frac{1}{2\pi s_2^2} \right)^{n/2} \cdot e^{-(m+n)/2} \quad \dots(16-42)$$

In  $\Theta_0$ , we have  $\mu_1 = \mu_2 = \mu$  and the likelihood function is given by :

$$L(\Theta_0) = \left( \frac{1}{2\pi\sigma_1^2} \right)^{m/2} \cdot \exp \left[ -\frac{1}{2\sigma_1^2} \sum_{i=1}^m (x_{1i} - \mu)^2 \right] \\ \times \left( \frac{1}{2\pi\sigma_2^2} \right)^{n/2} \cdot \exp \left[ -\frac{1}{2\sigma_2^2} \sum_{j=1}^n (x_{2j} - \mu)^2 \right]$$

To obtain the maximum value of  $L(\Theta_0)$  for variations in  $\mu$ ,  $\sigma_1^2$  and  $\sigma_2^2$ , it will be seen that estimate of  $\mu$  is obtained as the root of a cubic equation

$$\cdot \frac{m^2(\bar{x}_1 - \mu)}{\sum_{i=1}^m (x_{1i} - \hat{\mu})^2} + \frac{n^2(\bar{x}_2 - \mu)}{\sum_{j=1}^n (x_{2j} - \hat{\mu})^2} \quad \dots(16-43)$$

and is thus a complicated function of the sample observations. Consequently the likelihood ratio criterion  $\lambda$  will be a complex function of the observations and

its distribution is quite tedious since it involves the ratio of two variances. Consequently, it is impossible to obtain the critical region  $0 < \lambda < \lambda_0$ , for given  $\alpha$ , since the distribution of the population variances is ordinarily unknown. However, in any given instance the cubic equation (16-43) can be solved for  $\mu$  by numerical analysis technique and thus  $\lambda$  can be computed. Finally, as an approximate test,  $-2 \log_e \lambda$  can be regarded as a  $\chi^2$ -variate with 1 d.f. (c.f. Theorem 16-2).

**Case 2. Population Variances are equal, i.e.,  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , (say).**  
In this case

$$\Theta = \{(\mu_1, \mu_2, \sigma^2) : -\infty < \mu_i < \infty, \sigma^2 > 0, (i = 1, 2)\}$$

$$\Theta_0 = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$$

The likelihood function is then given by

$$L = \left( \frac{1}{2\pi\sigma^2} \right)^{(m+n)/2} \cdot \exp \left[ -\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^m (x_{1i} - \mu_1)^2 + \sum_{j=1}^n (x_{2j} - \mu_2)^2 \right\} \right] \quad \dots(16-44)$$

For  $\mu_1, \mu_2, \sigma^2 \in \Theta$ , the maximum likelihood equations are given by

$$\begin{aligned} \frac{\partial}{\partial \mu_1} \log L = 0 &\Rightarrow \hat{\mu}_1 = \bar{x}_1 \\ \frac{\partial}{\partial \mu_2} \log L = 0 &\Rightarrow \hat{\mu}_2 = \bar{x}_2 \end{aligned} \quad \dots(16-45)$$

$$\text{and } \frac{\partial}{\partial \sigma^2} \log L = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{m+n} [\sum (x_{1i} - \hat{\mu}_1)^2 + \sum (x_{2j} - \hat{\mu}_2)^2]$$

$$\begin{aligned} \Rightarrow \hat{\sigma}^2 &= \frac{1}{m+n} [\sum (x_{1i} - \bar{x}_1)^2 + \sum (x_{2j} - \bar{x}_2)^2] \\ &= \frac{1}{m+n} [ms_1^2 + ns_2^2] \end{aligned} \quad \dots(16-45a)$$

Substituting the values from (16-45) and (16-45a) in (16-44), we get

$$L(\hat{\Theta}) = \left[ \frac{(m+n)}{2\pi(ms_1^2 + ns_2^2)} \right]^{(m+n)/2} \cdot \exp \left[ -\frac{1}{2}(m+n) \right] \quad \dots(16-46)$$

In  $\Theta_0$ ,  $\mu_1 = \mu_2 = \mu$ , (say), and we get

$$\begin{aligned} L(\Theta_0) &= \left( \frac{1}{2\pi\sigma^2} \right)^{(m+n)/2} \cdot \exp \left[ -\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^m (x_{1i} - \mu)^2 \right. \right. \\ &\quad \left. \left. + \sum_{j=1}^n (x_{2j} - \mu)^2 \right\} \right] \end{aligned} \quad \dots(16-47)$$

$$\Rightarrow \log L(\Theta_0) \approx C - \frac{m+n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \left[ \sum_i (x_{1i} - \mu)^2 + \sum_j (x_{2j} - \mu)^2 \right],$$

where  $C$  is a constant independent of  $\mu$  and  $\sigma^2$ . The likelihood equation for estimating  $\mu$  gives

$$\begin{aligned}\frac{\partial}{\partial \mu} \log L &= \frac{1}{\sigma^2} \left[ \sum_{i=1}^m (x_{1i} - \mu) + \sum_{j=1}^n (x_{2j} - \mu) \right] = 0 \\ \Rightarrow \quad & (m\bar{x}_1 + n\bar{x}_2) - (m+n)\mu = 0 \\ \Rightarrow \quad & \hat{\mu} = \frac{1}{m+n} [m\bar{x}_1 + n\bar{x}_2] \quad \dots(16-48)\end{aligned}$$

Also  $\frac{\partial^2}{\partial \sigma^2} \log L = 0$

$$\begin{aligned}\Rightarrow \quad & -\frac{(m+n)}{2\sigma^2} + \frac{1}{2\sigma^4} [\sum(x_{1i} - \mu)^2 + \sum(x_{2j} - \mu)^2] = 0 \\ \Rightarrow \quad & \hat{\sigma}^2 = \frac{1}{m+n} [\sum(x_{1i} - \hat{\mu})^2 + \sum(x_{2j} - \hat{\mu})^2] \quad \dots(16-49)\end{aligned}$$

But  $\sum_{i=1}^m (x_{1i} - \hat{\mu})^2 = \sum_{i=1}^m (x_{1i} - \bar{x}_1 + \bar{x}_1 - \hat{\mu})^2$

$$= \sum(x_{1i} - \bar{x}_1)^2 + m(\bar{x}_1 - \hat{\mu})^2,$$

the product term vanishes since

$$\begin{aligned}\sum_i (x_{1i} - \bar{x}_1) &\approx 0 \\ \therefore \sum_{i=1}^m (x_{1i} - \hat{\mu})^2 &= ms_1^2 + m \left[ \bar{x}_1 - \frac{m\bar{x}_1 + n\bar{x}_2}{m+n} \right]^2 \\ &= ms_1^2 + \frac{mn^2(\bar{x}_1 - \bar{x}_2)^2}{(m+n)^2}\end{aligned}$$

Similarly, we shall get

$$\sum_{j=1}^n (x_{2j} - \hat{\mu})^2 = ns_2^2 + \frac{nm^2(\bar{x}_2 - \bar{x}_1)^2}{(m+n)^2}$$

Substituting in (16-49), we get

$$\hat{\sigma}^2 = \frac{1}{m+n} \left[ ms_1^2 + ns_2^2 + \frac{mn}{m+n} (\bar{x}_1 - \bar{x}_2)^2 \right] \quad \dots(16-49a)$$

Substituting from (16-48) and (16-49a) in (16-47), we get

$$L(\hat{\Theta}_0) = \left\{ \frac{(m+n)}{2\pi \left( ms_1^2 + ns_2^2 + \frac{mn}{m+n} (\bar{x}_1 - \bar{x}_2)^2 \right)} \right\}^{(m+n)/2} \times \exp \left( -\frac{m+n}{2} \right) \quad \dots(16-50)$$

$$\therefore \lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left\{ \frac{ms_1^2 + ns_2^2}{ms_1^2 + ns_2^2 + \frac{mn}{m+n} (\bar{x}_1 - \bar{x}_2)^2} \right\}^{(m+n)/2}$$

$$= \left\{ \frac{1}{\left( 1 + \frac{mn(\bar{x}_1 - \bar{x}_2)^2}{(m+n)(ms_1^2 + ns_2^2)} \right)} \right\}^{(m+n)/2} \quad \dots (16.51)$$

We know that (c.f. § 14-2-10), under the null hypothesis  $H_0 : \mu_1 = \mu_2$ , the statistic

$$t = -\frac{\bar{x}_1 - \bar{x}_2}{S \sqrt{\frac{1}{m} + \frac{1}{n}}}, \quad \dots (16.52)$$

where  $S^2 = \frac{1}{m+n-2}(ms_1^2 + ns_2^2) \quad \dots (16.52a)$

follows Student's  $t$ -distribution with  $(m+n-2)$  d.f. Thus in terms of  $t$ , we get

$$\lambda = \left[ 1 + \frac{t^2}{m+n-2} \right]^{-(m+n)/2} \quad \dots (16.53)$$

As in § 16-7-1, the test can as well be carried with  $t$  rather than with  $\lambda$ . The critical region  $0 < \lambda < \lambda_0$ , transforms to the critical region of the type

$$t^2 > (m+n-2) \left[ \frac{1}{\lambda_0^2/(m+n)} - 1 \right] = A^2, \text{ (say)}$$

i.e., by

$$|t| > A,$$

where  $A$  is determined so that

$$P [|t| > A | H_0] = \alpha \quad \dots (16.55)$$

Since under  $H_0$ , the statistic  $t$  follows Student's  $t$ -distribution with  $(m+n-2)$  d.f., we get from (16.55)

$$A = t_{m+n-2}(\alpha/2) \quad \dots (16.56)$$

where,  $t_n(\alpha)$  is the right  $100\alpha\%$  point of the  $t$ -distribution with  $n$  d.f.

Thus for testing the null hypothesis

$$H_0 : \mu_1 = \mu_2 ; \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$$

against the alternative

$$H_1 : \mu_1 \neq \mu_2, \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0,$$

we have the two-tailed  $t$ -test defined as follows :

If  $|t| = \left| \frac{\bar{x}_1 - \bar{x}_2}{S \sqrt{\frac{1}{m} + \frac{1}{n}}} \right| > t_{m+n-2}(\alpha/2)$

reject  $H_0$ , otherwise  $H_0$  may be accepted.

**Remarks.** 1. Proceeding similarly as in Remarks to § 16-7-1, we can obtain the critical regions for testing

$$H_0 : \mu_1 = \mu_2 ; \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$$

against the alternative hypothesis

$$H_1 : \mu_1 > \mu_2 ; \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$$

$$\text{or} \quad H_1' : \mu_1 < \mu_2 ; \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$$

We give below, in a tabular form the critical region, the test statistic and the confidence interval for testing the hypothesis

$$H_0 : \delta = \mu_1 - \mu_2 = \delta_0, (\text{say}),$$

against various alternatives, viz.,  $\delta > \delta_0$ ,  $\delta < \delta_0$  or  $\delta \neq \delta_0$ .

2. For testing  $H_0 : \delta = \delta_0$  against the alternative  $H_1 : \delta < \delta_0$ , the roles of  $x_1$  and  $x_2$  are interchanged and the case 1 of the table is applied.

3. If  $\delta_0 = 0$ , the above test reduces to testing  $H_0 : \mu_1 = \mu_2$ , i.e., the equality of two population means.

4. If the two population variances are not equal, then for testing  $H_0 : \delta = \delta_0$ , we use Fisher-Behrens' d-test.

S. No.	Alternative Hypothesis	Test	Test statistic	Reject $H_0$ at level of significance $\alpha$ if	$(1 - \alpha)$ confi- dence interval of $\delta$
1.	$\delta > \delta_0$	Right tailed	$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{S \sqrt{\frac{1}{m} + \frac{1}{n}}}$	$ t  > t_{m+n-2}(\alpha) = t_1, (\text{say})$	$\delta \geq (\bar{x}_1 - \bar{x}_2) - t_1 S \sqrt{\frac{1}{m} + \frac{1}{n}}$
2.	$\delta \neq \delta_0$	Two tailed	—do—	$ t  > t_{m+n-2}(\alpha/2) = t_2, (\text{say})$	$(\bar{x}_1 - \bar{x}_2) - t_2 S \sqrt{\frac{1}{m} + \frac{1}{n}} \leq \delta \leq (\bar{x}_1 - \bar{x}_2) + t_2 S \sqrt{\frac{1}{m} + \frac{1}{n}}$

**16.7.3. Test for the Equality of Means of Several Normal Populations.** Let  $X_{ij}$ , ( $j = 1, 2, \dots, n_i$ ;  $i = 1, 2, \dots, k$ ) be  $k$  independent random samples from  $k$  normal populations with means  $\mu_1, \mu_2, \dots, \mu_k$  respectively and unknown but common variance  $\sigma^2$ . In other words, the  $k$  normal populations are supposed to be *homoscedastic*. We want to test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu \quad (\text{say}), \text{(unspecified)}$$

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2 \quad (\text{say}), \text{(unspecified)}$$

against the alternative hypothesis

$$H_1 : \mu_i \text{'s are not all equal},$$

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2, \text{(unspecified)}$$

Thus we have

$$\Theta = \{(\mu_1, \mu_2, \dots, \mu_k, \sigma^2) : -\infty < \mu_i < \infty, (i = 1, 2, \dots, k); \sigma^2 > 0\}$$

and  $\Theta_0 = \{(\mu_1, \mu_2, \dots, \mu_k, \sigma^2) : -\infty < \mu_i = \mu < \infty, (i = 1, 2, \dots, k); \sigma^2 > 0\}$

The likelihood function of the sample observations is given by

$$L(\Theta) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2 \right] \quad \dots(16.57)$$

where  $n = \sum_{i=1}^k n_i$ .

For variations of  $\mu_i$ , ( $i = 1, 2, \dots, k$ ) and  $\sigma^2$  in  $\Theta$ , the maximum likelihood estimates are given by

$$\begin{aligned} \frac{\partial}{\partial \mu_i} \log L(\Theta) &= 0 \Rightarrow \sum_j (x_{ij} - \mu_i) = 0 \\ \Rightarrow \hat{\mu}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} = \bar{x}_i && \dots(16.58) \\ \frac{\partial}{\partial \sigma^2} \log L(\Theta) &= 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_i \sum_j (x_{ij} - \hat{\mu}_i)^2 \\ \Rightarrow \hat{\sigma}^2 &= \frac{1}{n} \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = \frac{S_W}{n}, \text{ (say),} && \dots(16.58a) \end{aligned}$$

where in ANOVA (Analysis of Variance) terminology,  $S_W$  is called *within sample sum of squares (W.S.S.)*.

In  $\Theta_0$ , the only variable parameters are  $\mu$  and  $\sigma^2$  and we have

$$L(\Theta_0) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_i \sum_j (x_{ij} - \mu)^2 \right\} \quad \dots(16.59)$$

The MLE's of  $\mu$  and  $\sigma^2$  are given by

$$\begin{aligned} \frac{\partial}{\partial \mu} \log L(\Theta_0) &= 0 \Rightarrow \sum_i \sum_j (x_{ij} - \mu) = 0 \\ \Rightarrow \hat{\mu} &= \frac{1}{n} \sum_i \sum_j x_{ij} = \bar{x} && \dots(16.60) \end{aligned}$$

and  $\frac{\partial}{\partial \sigma^2} \log L(\Theta_0) = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_i \sum_j (x_{ij} - \hat{\mu})^2$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_i \sum_j (x_{ij} - \bar{x})^2 = \frac{S_T}{n}, \text{ (say),} \quad \dots(16.60a)$$

where in ANOVA terminology,  $S_T$ , is called *total sum of squares (T.S.S.)*

Substituting from (16.58) and (16.58a) in (16.57) and from (16.60) and (16.60a) in (16.59), we get respectively

$$L(\hat{\Theta}) = \left( \frac{n}{2\pi S_W} \right)^{n/2} \cdot \exp \left( -\frac{n}{2} \right) \quad \dots(16-61)$$

and  $L(\hat{\Theta}_0) = \left( \frac{n}{2\pi S_T} \right)^{n/2} \cdot \exp \left( -\frac{n}{2} \right) \quad \dots(16-62)$

$$\therefore \lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left( \frac{S_W}{S_T} \right)^{n/2} \quad \dots(16-63)$$

We have

$$\begin{aligned} S_T &= \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 + 2 \sum_i [(\bar{x}_i - \bar{x}) \sum_j (x_{ij} - \bar{x}_i)] \end{aligned}$$

But  $\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = 0$ , being the algebraic sum of the deviations of the observations of the  $i$ th sample from its mean.

$$\begin{aligned} \therefore S_T &= \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_i n_i (\bar{x}_i - \bar{x})^2 \\ &= S_W + S_B, \text{ (say)} \end{aligned} \quad \dots(16-63a)$$

where  $S_B = \sum_i n_i (\bar{x}_i - \bar{x})^2$ , in ANOVA terminology is called *between samples sum of squares (B.S.S.)*:

Substituting in (16-63), we get

$$\begin{aligned} \lambda &= \left( \frac{S_W}{S_W + S_B} \right)^{n/2} \\ &= \frac{1}{\left[ 1 + \frac{S_B}{S_W} \right]^{n/2}} \end{aligned} \quad \dots(16-64)$$

We know that under  $H_0$ , the statistic

$$F = \frac{S_B/(k-1)}{S_W/(n-k)} \quad \dots(16-65)$$

follows  $F$ -distribution with  $(k-1; n-k)$  d.f.

Substituting in (16-64), the likelihood ratio criterion  $\lambda$  in terms of  $F$  is given by

$$\lambda = \left[ 1 + \frac{k-1}{n-k} F \right]^{-n/2} \quad \dots(16-66)$$

Since  $\lambda$  is a monotonic function of  $F$ , the test can well be carried on with  $F$  as test statistic rather than with  $\lambda$ . The critical region for testing  $H_0$  against  $H_1$ , viz.,  $0 < \lambda < \lambda_0$ , is equivalently given by

$$\left[ 1 + \frac{k-1}{n-k} F \right]^{n/2} > \lambda_0^{-1}$$

$$\Rightarrow F > \frac{n-k}{k-1} [(\lambda_0)^{-2/n} - 1] = A, \text{ (say)}, \quad \dots(16-67)$$

where  $A$  is determined from the equation

$$P[F > A | H_0] = \alpha \quad \dots(16-67a)$$

Since  $F$  follows  $F$ -distribution with  $(k-1, n-k)$  d.f., we get

$$A = F_{k-1, n-k}(\alpha)$$

where  $F_{k-1, n-k}(\alpha)$  denotes the upper  $\alpha$ -point of the  $F$ -distribution with  $(k-1, n-k)$  d.f.

*Hence the test for testing*

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu, \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2 > 0$$

against the alternative hypothesis

$$H_1 : \mu_i's \text{ are not all equal}, \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2 > 0$$

is defined as follows :

Reject  $H_0$  if  $F > F_{k-1, n-k}(\alpha)$ , otherwise  $H_0$  may be accepted, where  $F$  is defined in (16-65).

**Remark.** In ANOVA terminology,  $S_B/(k-1)$  is called Between Samples Mean Sum of Squares (M.S.S.) while  $S_W/(n-k)$  is called Within Samples (or Error) Mean Sum of Squares and thus  $F$  is defined as

$$F = \frac{\text{Between Samples M.S.S.}}{\text{Within Samples M.S.S.}} \quad \dots(16-67c)$$

**16-7-4. Test for the Variance of a Normal Population.** Let us now consider the problem of testing if the variance of a normal population has a specified value  $\sigma_0^2$ , on the basis of a random sample  $x_1, x_2, \dots, x_n$  of size  $n$  from normal population  $N(\mu, \sigma^2)$ .

We want to test the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2, \text{ (specified),}$$

against the alternative hypothesis

$$H_1 : \sigma^2 \neq \sigma_0^2$$

Here we have

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$$

and  $\Theta_0 = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 = \sigma_0^2\}$

The likelihood function of the sample observations is given by

$$L = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \quad \dots(16-68)$$

As in § (16-7-1), [c.f. (16-27)], we shall get

$$L(\hat{\Theta}) = \left( \frac{1}{2\pi s^2} \right)^{n/2} \exp \left( -\frac{n}{2} \right) \quad \dots(16-69)$$

In  $\Theta_0$ , we have only one variable parameter, viz.,  $\mu$  and

$$L(\Theta_0) = \left( \frac{1}{2\pi\sigma_0^2} \right)^{n/2} \exp \left[ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2 \right] \quad \dots(16-70)$$

The MLE for  $\mu$  is given by

$$\frac{\partial}{\partial \mu} \log L = 0 \Rightarrow \hat{\mu} = \bar{x}$$

$$\therefore L(\hat{\Theta}_0) = \left( \frac{1}{2\pi\sigma_0^2} \right)^{n/2} \exp \left[ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

$$= \left( \frac{1}{2\pi\sigma_0^2} \right)^{n/2} \exp \left[ -\frac{ns^2}{2\sigma_0^2} \right] \quad \dots(16-71)$$

The likelihood ratio criterion is given by

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left[ \frac{s^2}{\sigma_0^2} \right]^{n/2} \exp \left[ -\frac{1}{2} \left( \frac{ns^2}{\sigma_0^2} - n \right) \right]$$

We know that under  $H_0$ , the statistic

$$\chi^2 = \frac{ns^2}{\sigma_0^2} \quad \dots(16-72)$$

follows chi-square distribution with  $(n - 1)$  d.f. In terms of  $\chi^2$ , we have

$$\lambda = \left[ \frac{\chi^2}{n} \right]^{n/2} \exp \left[ -\frac{1}{2} (\chi^2 - n) \right] \quad \dots(16-73)$$

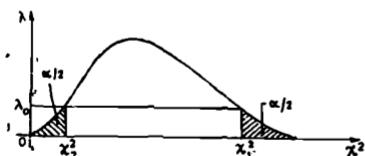
Since  $\lambda$  is a monotonic function of  $\chi^2$ , the test may be done using  $\chi^2$  as a criterion. The critical region  $0 < \lambda < \lambda_0$  is now equivalent to

$$(\chi^2/n)^{n/2} \exp \left[ -\frac{1}{2} (\chi^2 - n) \right] < \lambda_0$$

$$\text{or } \exp \left( -\frac{1}{2} \chi^2 \right) (\chi^2)^{n/2} < \lambda_0 \cdot (ne^{-1})^{n/2} = B, \text{ (say).} \quad \dots(16-74)$$

Since  $\chi^2$  has chi-square distribution with  $(n - 1)$  d.f., the critical region (16-74) is determined by a pair of intervals  $0 < \chi^2 < \chi_2^2$  and  $\chi_1^2 < \chi^2 < \infty$ , where  $\chi_1^2$  and  $\chi_2^2$  are to be determined such that the ordinates of (16-73) are equal, i.e.,

$$(\chi_1^2)^{n/2} \exp \left( -\frac{1}{2} \chi_1^2 \right) \\ = (\chi_2^2)^{n/2} \exp \left( -\frac{1}{2} \chi_2^2 \right)$$



Critical region is shown as shaded region in the above diagram.

In other words,  $\chi_1^2$  and  $\chi_2^2$  are defined by the equations

$$\left. \begin{aligned} P(\chi^2 > \chi_1^2) &= \alpha/2 \\ P(\chi^2 > \chi_2^2) &= 1 - \frac{\alpha}{2} \end{aligned} \right\} \quad \dots(16-75)$$

and

In other words,

$$\chi_1^2 = \chi_{n-1}^2(\alpha/2) \text{ and } \chi_2^2 = \chi_{n-1}^2(1 - \alpha/2),$$

where  $\chi_{n-1}^2(\alpha)$  is the upper  $\alpha$ -point of the chi-square distribution with  $(n-1)$  d.f. Thus the critical region for testing  $H_0: \sigma^2 = \sigma_0^2$  against  $H_1: \sigma^2 \neq \sigma_0^2$ , is a two-tailed region given by

$$\chi^2 > \chi_{n-1}^2(\alpha/2) \text{ and } \chi^2 < \chi_{n-1}^2(1 - \alpha/2) \quad \dots(16-76)$$

Thus, in this case we have a two-tailed test.

**Remarks.** 1. If we want to test  $H_0: \sigma^2 = \sigma_0^2$  against the alternative hypothesis  $H_1: \sigma^2 < \sigma_0^2$  we get a one-tailed (left-tailed) test with critical region  $\chi^2 < \chi_{n-1}^2(1 - \alpha)$  while for testing  $H_0$  against  $H_1: \sigma^2 > \sigma_0^2$ , we have a right tailed test with critical region  $\chi^2 > \chi_{n-1}^2(\alpha)$ .

We give below in a tabular form, the test statistic, the test criterion and the confidence interval for the parameter for testing  $H_0: \sigma^2 = \sigma_0^2, \mu$  (unknown), against various alternative hypotheses.

NORMAL POPULATION  $N(\mu, \sigma^2)$ ;  $\mu$  UNKNOWN;  $H_0: \sigma^2 = \sigma_0^2$

S. No.	Alternative hypothesis	Test	Test statistic	Reject $H_0$ at ' $\alpha$ ' level of significance if	$(1 - \alpha)$ confidence interval for $\sigma^2$
1.	$\sigma^2 > \sigma_0^2$	Right-tailed test	$\chi^2 = \frac{ns^2}{\sigma_0^2}$	$\chi^2 > \chi_{n-1}^2(\alpha)$	$\sigma^2 \geq \frac{ns^2}{\chi_{n-1}^2(\alpha)}$
2.	$\sigma^2 < \sigma_0^2$	Left-tailed test	— do —	$\chi^2 < \chi_{n-1}^2(1 - \alpha)$	$\sigma^2 \leq \frac{ns^2}{\chi_{n-1}^2(1 - \alpha)}$
3.	$\sigma^2 \neq \sigma_0^2$	Two-tailed test	— do —	$\chi^2 > \chi_{n-1}^2(\alpha/2)$ and $\chi^2 < \chi_{n-1}^2(1 - \alpha/2)$	$\frac{ns^2}{\chi_{n-1}^2(\alpha/2)} \leq \sigma^2 \leq \frac{ns^2}{\chi_{n-1}^2(1 - \alpha/2)}$

2. If we want to test the null hypothesis  $H_0: \sigma^2 = \sigma_0^2$  against the various alternative hypotheses, viz.,  $\sigma^2 \geq \sigma_0^2$  or  $\sigma^2 < \sigma_0^2$  or  $\sigma^2 \neq \sigma_0^2$  for the normal population  $N(\mu, \sigma^2)$ , where  $\mu$  is known then the test statistic, the critical region and the confidence interval for  $\sigma^2$  can be obtained from the table given above on replacing  $(n-1)$  by  $n$  and  $ns^2$  by the expression  $\sum_{i=1}^n (x_i - \mu)^2$ .

**16.7.5. Test for Equality of Variances of two Normal Populations.** Consider two normal populations  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$  where the means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2, \sigma_2^2$  are unspecified. We want to test the hypothesis:

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$  (unspecified), with  $\mu_1$  and  $\mu_2$  (unspecified) against the alternative hypothesis

$$H_1 : \sigma_1^2 \neq \sigma_2^2 ; \mu_1 \text{ and } \mu_2 \text{ (unspecified).}$$

If  $x_{1i}, (i = 1, 2, \dots, m)$  and  $x_{2j}, (j = 1, 2, \dots, n)$  be independent random samples of sizes  $m$  and  $n$  from  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$  respectively then

$$L = \left( \frac{1}{2\pi\sigma_1^2} \right)^{m/2} \exp \left[ -\frac{1}{2\sigma_1^2} \sum_{i=1}^m (x_{1i} - \mu_1)^2 \right] \\ \times \left( \frac{1}{2\pi\sigma_2^2} \right)^{n/2} \exp \left[ -\frac{1}{2\sigma_2^2} \sum_{j=1}^n (x_{2j} - \mu_2)^2 \right] \quad \dots(16.77)$$

In this case

$$\Theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 : -\infty < \mu_i < \infty; \sigma_i^2 > 0, (i = 1, 2)\}$$

$$\text{and } \Theta_0 = \{(\mu_1, \mu_2, \sigma^2) : -\infty < \mu_i < \infty; (i = 1, 2), \sigma^2 > 0\}$$

As in § 16.7.2 [c.f. (16.42)],

$$L(\hat{\Theta}) = \left( \frac{1}{2\pi s_1^2} \right)^{m/2} \cdot \left( \frac{1}{2\pi s_2^2} \right)^{n/2} \cdot \exp \left[ -\frac{1}{2} (m + n) \right] \quad \dots(16.78)$$

where  $s_1^2$  and  $s_2^2$  are as defined in (16.41a).

In  $\Theta_0$ , the likelihood function (16.77) is given by

$$L(\Theta_0) = \left[ \frac{1}{2\pi\sigma^2} \right]^{(m+n)/2} \cdot \exp \left[ -\frac{1}{2\sigma^2} \left\{ \sum_i (x_{1i} - \mu_1)^2 + \sum_j (x_{2j} - \mu_2)^2 \right\} \right] \quad \dots(16.79)$$

and the MLE's for  $\mu_1, \mu_2$  and  $\sigma^2$  are now given by

$$\hat{\mu}_1 = \bar{x}_1, \hat{\mu}_2 = \bar{x}_2 \quad \dots(16.80)$$

$$\text{and } \hat{\sigma}^2 = \frac{1}{(m+n)} \left[ \sum_i (x_{1i} - \hat{\mu}_1)^2 + \sum_j (x_{2j} - \hat{\mu}_2)^2 \right] \\ = \frac{1}{m+n} \left[ \sum_i (x_{1i} - \bar{x}_1)^2 + \sum_j (x_{2j} - \bar{x}_2)^2 \right] \\ = \frac{ms_1^2 + ns_2^2}{m+n} \quad \dots(16.80a)$$

Substituting from (16.80) and (16.80a) in (16.79), we get

$$L(\hat{\Theta}_0) = \left[ \frac{m+n}{2\pi(ms_1^2 + ns_2^2)} \right]^{(m+n)/2} \cdot \exp \left[ -\frac{1}{2} (m+n) \right] \quad \dots(16.81)$$

$$\begin{aligned} \therefore \lambda &= \frac{\hat{L}(\hat{\Theta}_0)}{\hat{L}(\hat{\Theta})} \\ &= (m+n)^{(m+n)/2} \left\{ \frac{(s_1^2)^{m/2} (s_2^2)^{n/2}}{[ms_1^2 + ns_2^2]^{(m+n)/2}} \right\} \\ &= \frac{(m+n)^{(m+n)/2}}{m^{m/2} \cdot n^{n/2}} \left\{ \frac{(ms_1^2)^{m/2} (ns_2^2)^{n/2}}{[ms_1^2 + ns_2^2]^{(m+n)/2}} \right\} \quad \dots(16-82) \end{aligned}$$

We know that under  $H_0$ , the statistic

$$F = \frac{\sum(x_{1i} - \bar{x}_1)^2/(m-1)}{\sum(x_{2j} - \bar{x}_2)^2/(n-1)} = \frac{s_1^2}{s_2^2}, \quad \dots(16-83)$$

follows  $F$ -distribution with  $(m-1, n-1)$  d.f. (16-83) also implies

$$\begin{aligned} F &= \frac{m(n-1)s_1^2}{n(m-1)s_2^2} \\ \Rightarrow \left( \frac{m-1}{n-1} \right) F &= \frac{ms_1^2}{ns_2^2} \quad \dots(16-83a) \end{aligned}$$

Substituting in (16-82) and simplifying, we get

$$\lambda = \frac{(m+n)^{(m+n)/2}}{m^{m/2} n^{n/2}} \left\{ \frac{\left( \frac{m-1}{n-1} F \right)^{m/2}}{1 + \frac{m-1}{n-1} F} \right\}^{(m+n)/2} \quad \dots(16-84)$$

Thus  $\lambda$  is a monotonic function of  $F$  and hence the test can be carried on with  $F$ , defined in (16-83) as test statistic. The critical region  $0 < \lambda < \lambda_0$  can be equivalently seen to be given by pair of intervals  $F \leq F_1$  and  $F \geq F_2$ , where  $F_1$  and  $F_2$  are determined so that under  $H_0$

$$P(F \geq F_2) = \alpha/2 \text{ and } P(F \leq F_1) = 1 - \alpha/2$$

Since, under  $H_0$ ,  $F$  follows Snedecor's  $F$ -distribution with  $(m-1, n-1)$  d.f., we have

$$F_2 = F_{m-1, n-1}(\alpha/2) \text{ and } F_1 = F_{m-1, n-1}(1 - \alpha/2),$$

where  $F_{m, n}(\alpha)$  is the upper  $\alpha$ -point of  $F$ -distribution with  $(m, n)$  d.f. Consequently for testing  $H_0: \sigma_1^2 = \sigma_2^2$  against the alternative hypothesis  $H_1: \sigma_1^2 \neq \sigma_2^2$ , we have a two-tailed  $F$ -test, the *critical region* being given by

$$F > F_{m-1, n-1}(\alpha/2) \text{ and } F < F_{m-1, n-1}(1 - \alpha/2) \quad \dots(16-85)$$

where  $F$  is defined in (16-83) or (16-83a).

**Remark.** Let us suppose that we want to test the hypothesis

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = \delta_0^2$$

Without loss of generality, we can assume that  $S_1^2 > S_2^2$ , where  $S_1^2$  and  $S_2^2$  are unbiased estimates of  $\sigma_1^2$  and  $\sigma_2^2$  respectively. We know that the statistic

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \cdot \frac{1}{\delta_0^2}, \text{ (under } H_0).$$

follows  $F$ -distribution with  $(m - 1, n - 1)$  d.f. The test-statistic, the test criterion and  $(1 - \alpha)$  confidence interval for the parameter for various alternative hypotheses are given in the following table.

If  $\delta_0 = 1$ , the above test reduces to testing the equality of population variances.

$$\text{NORMAL POPULATION: } H_0: \frac{\sigma_1^2}{\sigma_2^2} = \delta_0^2$$

<i>S. No.</i>	<i>Alternative Hypothesis</i>	<i>Test</i>	<i>Test Statistic</i>	<i>Critical region at level of significance '<math>\alpha</math>'</i>	<i><math>(1 - \alpha)</math> confidence interval for <math>\frac{\sigma_1^2}{\sigma_2^2}</math></i>
1.	$\frac{\sigma_1^2}{\sigma_2^2} > \delta_0^2$	Right-tailed	$F = \frac{S_1^2}{S_2^2} \cdot \frac{1}{\delta_0^2}$	$F > F_{m-1, n-1}(\alpha)$	$\frac{\sigma_1^2}{\sigma_2^2} \geq \frac{S_1^2}{S_2^2} \times \frac{1}{F_{m-1, n-1}(\alpha)}$
2.	$\frac{\sigma_1^2}{\sigma_2^2} < \delta_0^2$	Left-tailed	— do —	$F < F_{m-1, n-1}(1 - \alpha)$	$\frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} \times \frac{1}{F_{m-1, n-1}(1 - \alpha)}$
3.	$\frac{\sigma_1^2}{\sigma_2^2} \neq \delta_0^2$	Two-tailed	— do —	$F > F_{m-1, n-1}(\alpha/2),$ and $F < F_{m-1, n-1}(1 - \alpha/2)$	$\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{m-1, n-1}(\alpha/2)} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{m-1, n-1}(1 - \alpha/2)}$

**16.7.6. Test for the Equality of Variances of Several Normal Populations.** Let  $X_{ij}$ , ( $j = 1, 2, \dots, n_i$ ) be a random sample of size  $n_i$  from the normal population  $N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, k$ . We want to test the null hypothesis :

$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$  (unspecified), with  $\mu_1, \mu_2, \dots, \mu_k$  (unspecified), against the alternative hypothesis :

$H_1 : \sigma_i^2$  ( $i = 1, 2, \dots, k$ ), are not all equal;  $\mu_1, \mu_2, \dots, \mu_k$  (unspecified).

Here we have

$$\Theta = \{\mu_1, \mu_2, \dots, \mu_k; \sigma_1^2, \sigma_2^2, \dots, \sigma_k^2\}: -\infty < \mu_i < \infty, \sigma_i^2 > 0, \quad (i = 1, 2, \dots, k)\}$$

$$\text{and } \Theta_0 = \{\mu_1, \mu_2, \dots, \mu_k; \sigma_1^2, \sigma_2^2, \dots, \sigma_k^2\}: -\infty < \mu_i < \infty, \sigma_i^2 = \sigma^2 > 0, \quad (i = 1, 2, \dots, k)\}$$

The likelihood function of the sample observations  $x_{ij}$ , ( $j = 1, 2, \dots, n_i$ ;  $i = 1, 2, \dots, k$ ) is given by :

$$L = \prod_{i=1}^k \left\{ \left( \frac{1}{2\pi\sigma_i^2} \right)^{n_i/2} \cdot \exp \left[ -\frac{1}{2\sigma_i^2} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2 \right] \right\} \quad \dots(16-86)$$

It can be easily seen that in  $\Theta$  the MLE's of  $\mu_i$ 's and  $\sigma_i$ 's are given by

$$\hat{\mu}_i = \bar{x}_i \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = s_i^2, \quad \dots(16-87)$$

$$\begin{aligned} \therefore L(\hat{\Theta}) &= \prod_{i=1}^k \left\{ \left( \frac{1}{2\pi s_i^2} \right)^{n_i/2} \cdot \exp \left[ -\frac{n_i}{2} \right] \right\} \\ &= \exp \left( -\frac{n}{2} \right) \cdot \prod_{i=1}^k \left[ \left( \frac{1}{2\pi s_i^2} \right)^{n_i/2} \right] \end{aligned} \quad \dots(16-88)$$

where  $n = \sum n_i$ .

In  $\Theta_0$ ,  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$  and therefore

$$L(\Theta_0) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left[ -\frac{1}{2\sigma^2} \sum_{j=1}^n (x_{ij} - \mu_i)^2 \right] \quad \dots(16-89)$$

The MLE's of  $\mu_i$ 's and  $\sigma^2$  are given by

$$\hat{\mu}_i = \bar{x}_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i,j} (x_{ij} - \bar{x}_i)^2 = \frac{1}{n} \sum_i n_i s_i^2 \quad \dots(16-90)$$

Substituting from (16-90) in (16-89), we get

$$\begin{aligned} L(\hat{\Theta}_0) &= \left( \frac{n}{2\pi \sum_i n_i s_i^2} \right)^{n/2} \cdot \exp \left( -\frac{n}{2} \right) \\ \therefore \lambda &= \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \frac{\frac{n^{n/2}}{\prod_{i=1}^k [(s_i^2)^{n_i/2}]}}{\left[ \sum_{i=1}^k n_i s_i^2 \right]^{n/2}} \\ &= \frac{\prod_{i=1}^k [(s_i^2)^{n_i/2}]}{(s^2)^{n/2}}, \text{ where } s^2 = \frac{1}{n} \sum n_i s_i^2 \\ &= \prod_{i=1}^k \left[ \left( \frac{s_i^2}{s^2} \right)^{n_i/2} \right] \end{aligned} \quad \dots(16-92)$$

$\lambda$  is thus a complicated function of sample observations and it is not easy to obtain its distribution. However, if  $n_i$ 's are large ( $i = 1, 2, \dots, k$ ), Theorem 16-2 provides an approximate test defined as follows :

For large  $n_i$ 's, the quantity  $-2 \log \lambda$  is approximately distributed as a chi-square variate with  $2k - (k + 1) = k - 1$  d.f.

The test can, however, be made even if  $n_i$ 's are not large. It has been investigated and found that the distribution of  $-2 \log \lambda$  is approximately a

$\chi^2$ -distribution with  $(k - 1)$  d.f. even for small  $n_i$ 's. However, a better approximation is provided by the Bartlett's test statistic :

$$\chi^2 = \frac{-2 \log \lambda'}{1 + \frac{1}{3(k-1)} \left[ \sum_i \left( \frac{1}{n_i} \right) - \frac{1}{\sum n_i} \right]}$$

where  $\lambda'$  is obtained from  $\lambda$  on replacing  $n_i$  by  $(n_i - 1)$  in (16.92), which follows  $\chi^2$ -distribution with  $(k - 1)$  d.f. Thus the test statistic, under  $H_0$  is given by

$$\chi^2 = \frac{\sum_{i=1}^k (n_i - 1) \log_e \left( \frac{s_i^2}{s^2} \right)}{1 + \frac{1}{3(k-1)} \left[ \sum_i \left( \frac{1}{n_i} \right) - \frac{1}{\sum n_i} \right]} \sim \chi^2_{k-1} \quad \dots (16.93)$$

The critical region for the test is, of course, the right-tail of the  $\chi^2$ -distribution given by

$$\chi^2 > \chi^2_{(k-1)}(\alpha), \quad \dots (16.94)$$

where  $\chi^2$  is defined in (16.93).

### EXERCISE 16 (b)

1. (a) Define 'Likelihood Ratio Test'. Under what circumstances would you recommend this test ?

(b) Let  $x_1, x_2, \dots, x_n$  be a random sample from a normal distribution  $N(\theta_1, \theta_2)$ . Use likelihood ratio test to obtain BCR of size  $\alpha$  under  $H_0 : \theta_1 = 0$  against  $H_1 : \theta_1 \neq 0$ .

2. (a) Let  $p_\theta(x)$  be the density of a random variable with the mixed second derivative  $\frac{\partial^2 \log p_\theta(x)}{\partial \theta \partial x} \geq 0$  for all  $x$  and  $\theta$ . Then show that the family has monotone likelihood ratio in  $x$ .

3. Discuss the general method of construction of likelihood ratio test. Consider  $n$  Bernoullian trials with probability of success  $p$  for each trial. Derive the likelihood-ratio test for testing  $H_0 : p = p_0$  against  $H_1 : p > p_0$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986]

4. Let  $X_1, X_2, \dots, X_n$  be a random sample from a Poisson distribution with parameter  $\theta$ . Derive the likelihood ratio test for  $H_0 : \theta = \theta_0$  against  $H_1 : \theta > \theta_0$ . Show that this is identical with the corresponding UMP test.

5. (a) Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal population with unknown mean  $\mu$  and, known variance  $\sigma^2$ . Develop the likelihood ratio test for testing  $H_0 : \mu = \mu_0$  (specified) against (i)  $H_1 : \mu > \mu_0$  and (ii)  $H_1 : \mu < \mu_0$ .

(b) Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$ , where  $\sigma^2$  is known. Develop the likelihood ratio test for testing  $H_0 : \mu = \mu_0$  (specified) against  $H_1 : \mu < \mu_0$ .

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

(c) Find by the method of likelihood ratio testing, a test for the null hypothesis  $H_0 : \theta = \theta_0$  for a normal  $(\mu, \sigma^2)$  population,  $\sigma^2$  known.

[*Calcutta Univ. B.Sc. (Maths Hons.), 1989*]

6. Discuss the general method of construction of likelihood ratio test.

Let  $X_1, X_2, \dots, X_n$  be a random sample from a  $N(\mu, \theta)$  population where  $\theta$  is the unknown variance and  $\mu$  is known. Obtain a likelihood ratio test for testing a simple  $H_0 : \theta = \theta_0$  against  $H_1 : \theta > \theta_0$ .

[*Delhi Univ. B.Sc. (Stat. Hons.), 1993*]

7. (a) Develop the likelihood ratio test for testing  $H_0 : \mu = \mu_0$  based on a random sample of size  $n$  from  $N(\mu, \sigma^2)$  population.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1982*]

(b) Let  $x_1, x_2, \dots, x_n$  be a random sample from a normal population with mean  $\mu$  and variance  $\sigma^2$ ,  $\mu$  and  $\sigma^2$  being unknown. We wish to test  $H_0 : \mu = \mu_0$  (specified) against  $H_1 : \mu \neq \mu_0, 0 < \sigma^2 < \infty$ .

Show that the Likelihood Ratio Test is same as the two tailed  $t$ -test.

[*Delhi Univ. M.A. (Eco.), 1986*]

(c) Describe the likelihood ratio test.

The random variable  $X$  follows normal distribution with mean  $\theta_1$  and variance  $\theta_2$ . The parameter space is

$$\Theta = \{(\theta_1, \theta_2) : -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\}.$$

$$\text{Let } \Theta_0 = \{(\theta_1, \theta_2) : \theta_1 = 0, 0 < \theta_2 < \infty\}.$$

Test the hypothesis  $H_0 : \theta_1 = 0, \theta_2 > 0$  against the alternative composite hypothesis  $H_1 : \theta_1 \neq 0, \theta_2 > 0$ .

[*Madras Univ. B.Sc., 1988*]

8. Discuss the general method of construction of likelihood ratio test. Given  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$ , where all the parameters  $\mu_1, \mu_2, \sigma_1^2$  and  $\sigma_2^2$  are unspecified, develop the LR test for testing  $H_0 : \sigma_1^2 = \sigma_2^2$  against  $H_1 : \sigma_1^2 \neq \sigma_2^2$ .

[*Delhi Univ. B.Sc. (Stat. Hons.), 1983*]

9. Describe likelihood ratio test and state its important properties.

Let  $\bar{X}_1$  and  $\bar{X}_2$  be  $N(\mu_1, \sigma^2)$  and  $N(\mu_2, \sigma^2)$  respectively where the means and variance are unspecified. Develop LR test for testing  $H_0 : \mu_1 = \mu_2$  against  $H_1 : \mu_1 \neq \mu_2$ .

*OR*

Construct LR test for testing  $H_0 : \theta = \theta_0$  against all its alternatives in  $N(\theta, \sigma^2)$ , where  $\sigma^2$  is known.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

10. Show that the likelihood ratio test for testing the equality of variances of two normal distributions is the usual  $F$ -test.

11. Show that the likelihood ratio test for testing  $H_0 : \alpha = 0$  against  $H_1 : \alpha \neq 0$ , based on a random sample of size  $n$  from

$$f(x; \alpha; \beta) = \frac{1}{2\beta} ; \alpha - \beta \leq x \leq \alpha + \beta$$

is  $(R/Z)^n$  where  $R = X_{(n)} - X_{(1)}$  and  $Z = \max [-X_{(1)}, X_{(n)}]$ .

[*Delhi B.Sc. (Stat. Hons.), 1989, 1985*]

12. Show that the likelihood ratio principle leads to the same test, when testing a simple hypothesis against an alternative simple hypothesis, as that given by Neyman-Pearson theorem. [*Madras Univ. B.Sc., 1988*]

**16.8. Non-parametric Methods.** Most of the statistical tests that we have discussed so far had the following two features in common.

(i) The form of the frequency function of the parent population from which the samples have been drawn is assumed to be known, and

(ii) They were concerned with testing statistical hypothesis about the parameters of this frequency function or estimating its parameters.

For example, almost all the exact (small) sample tests of significance are based on the fundamental assumption that the parent population is normal and are concerned with testing or estimating the means and variances of these populations. Such tests, which deal with the parameters of the population are known as *Parametric Tests*. Thus, a parametric statistical test is a test whose model specifies certain conditions about the parameters of the population from which the samples are drawn.

On the other hand, a *Non-parametric (N.P.) Test* is a test that does not depend on the particular form of the basic frequency function from which the samples are drawn. In other words, non-parametric test does not make any assumption regarding the form of the population.

However, certain assumptions associated with N.P. tests are :

(i) Sample observations are independent.

(ii) The variable under study is continuous.

(iii) p.d.f. is continuous.

(iv) Lower order moments exist.

Obviously these assumptions are fewer and much weaker than those associated with parametric tests.

**16.8.1. Advantages and Disadvantages of N.P. Methods over Parametric Methods.** Below we shall give briefly the comparative study of parametric and non-parametric methods and their relative merits and demerits.

**Advantages of N.P. Methods :**

(i) N.P. methods are readily comprehensible, very simple and easy to apply and do not require complicated sample theory.

(ii) No assumption is made about the form of the frequency function of the parent population from which sampling is done.

(iii) No parametric technique will apply to the data which are mere classification (*i.e.*, which are measured in nominal scale), while N.P. methods exist to deal with such data.

(iv) Since the socio-economic data are not, in general, normally distributed, N.P. tests have found applications in Psychometry, Sociology and Educational Statistics.

(v) N.P. tests are available to deal with the data which are given in ranks or whose seemingly numerical scores have the strength of ranks. For instance, no

parametric test can be applied if the scores are given in grades such as  $A^+, A^-, B, A, B^+$ , etc.

### Disadvantages of N.P. Tests.

(i) N.P. tests can be used only if the measurements are nominal or ordinal. Even in that case, if a parametric test exists it is more powerful than the N.P. test. In other words, if all the assumptions of a statistical model are satisfied by the data and if the measurements are of required strength, then the N.P. tests are wasteful of time and data.

(ii) So far, no N.P. methods exist for testing interactions in 'Analysis of Variance' model unless special assumptions about the additivity of the model are made.

(iii) N.P. tests are designed to test statistical hypothesis only and not for estimating the parameters.

**Remarks 1.** Since no assumption is made about the parent distribution, the N.P. methods are sometimes referred to as *Distribution Free* methods. These tests are based on the 'Order Statistic' theory. In these tests we shall be using median, range, quartile, inter-quartile range, etc., for which an ordered sample is desirable. By saying that  $x_1, x_2, \dots, x_n$  is an ordered sample we mean  $x_1 \leq x_2 \leq \dots \leq x_n$ .

2. The whole structure of the N.P. methods rests on a simple but fundamental property of order statistic, viz.

"*The distribution of the area under the density function between any two ordered observations is independent of the form of the density function*", which we shall now prove.

**16-8-2. Basic Distribution.** Let  $Z$  be a continuous random variable with a p.d.f.  $f(\cdot)$ . Let  $Z_1, Z_2, \dots, Z_n$  be a random sample of size  $n$  from  $f(\cdot)$  and let  $x_1, x_2, \dots, x_n$  be the corresponding ordered sample. Then the joint density of  $x_1, x_2, \dots, x_n$  is given by

$$g(x_1, x_2, \dots, x_n) = n! f(x_1) f(x_2) \dots f(x_n), -\infty < x_1 < x_2 < \dots < x_n < \infty \quad \dots(16.95)$$

the factor  $n!$  appearing since there are  $n!$  permutations of the sample observations and each gives rise to the same ordered sample.

Let us define

$$U_i = \int_{-\infty}^{x_i} f(z) dz = F(x_i), (i = 1, 2, \dots, n) \quad \dots(16.96)$$

where  $F(\cdot)$  is the distribution function of  $Z$ . But since  $F(x_i)$  is a uniform random variable on  $[0, 1]$ ,  $U_i$ , ( $i = 1, 2, \dots, n$ ), defined in (16.96) are random variables following uniform distribution on  $[0, 1]$ . Thus the joint density  $k(\cdot)$  of the random variables  $U_i$ , ( $i = 1, 2, \dots, n$ ) is given by

$$k(u_1, u_2, \dots, u_n) = n!, 0 \leq u_1 < u_2 < \dots < u_n \leq 1 \quad \dots(16.97)$$

and does not depend on  $f(\cdot)$ .

$$\begin{aligned} E(U_i) &= \int_0^1 \dots \int_0^{u_3} \int_0^{u_2} u_i n! du_1 du_2 \dots du_n \\ &= \frac{i}{n+1} \quad (\text{On simplification}) \quad \dots(16.98) \end{aligned}$$

Thus the expected area under  $f(\cdot)$  between two successive ordered observations is given by

$$E(U_i) - E(U_{i-1}) = \frac{i}{n+1} - \frac{i-1}{n+1} = \frac{1}{n+1}, \quad \dots(16.98a)$$

which is independent of  $f(\cdot)$ .

**16.8.3. Wald-Wolfowitz Run Test.** Suppose  $x_1, x_2, \dots, x_{n_1}$  is an ordered sample from a population with density  $f_1(\cdot)$  and let  $y_1, y_2, \dots, y_{n_2}$  be an independent ordered sample from another population with density  $f_2(\cdot)$ . We want to test if the samples have been drawn from the same population or from populations with the same density functions, i.e., if  $f_1(\cdot) = f_2(\cdot)$ .

Let us combine the two samples and arrange the observations in order of magnitude to give the combined ordered sample as, (say),

$$x_1 x_2 y_1 y_2 y_3 x_3 y_4 x_4 y_5 x_5 \dots \quad \dots(16.99)$$

**Run (Definition).** A run is defined as a sequence of letters of one kind surrounded by a sequence of letters of the other kind, and the number of elements in a run is usually referred to as the length ( $l$ ) of the run.

Thus in (16.99), we have in order, a run of  $x$  ( $l = 2$ ), a run of  $y$  ( $l = 3$ ), a run of  $x$  ( $l = 1$ ), a run of  $y$  ( $l = 1$ ) etc.

If both the samples come from the same population then there would be thorough mingling of  $x$ 's and  $y$ 's and consequently the number of runs in the combined sample would be large. On the other hand if the samples come from two different populations so that their ranges do not overlap, then there would be only two runs of the type  $x_1, x_2, \dots, x_{n_1}$  and  $y_1, y_2, \dots, y_{n_2}$ . Generally, any difference in mean and variance would tend to reduce the number of runs. Thus the alternative hypothesis will entail too few runs.

**Procedure.** In order to test the Null Hypothesis  $H_0 : f_1(\cdot) = f_2(\cdot)$  i.e., the samples have come from the same population we count the number of runs ' $U$ ' in the combined ordered sample.

Null hypothesis is rejected if  $U < u_0$ , where the value of  $u_0$  for given level of significance is determined from considering the distribution of  $U$  under  $H_0$ .

First of all let us find the probability of obtaining a specific arrangement (16.99) under  $H_0 : f_1(\cdot) = f_2(\cdot) = f(\cdot)$ , (say).

If  $X$ 's and  $Y$ 's are transformed to  $U$ 's and  $V$ 's by the relation :

$$U_i = \int_{-\infty}^{x_i} f(z) dz, \quad V_i = \int_{-\infty}^{y_i} f(z) dz,$$

then the joint p.d.f. of  $U$ 's and  $V$ 's becomes

$$g(u_1, u_2, \dots, u_n, v_1, v_2, \dots, v_n) = n_1! n_2! \quad \dots(16.100)$$

The probability of an arrangement (16.99) is obtained on integrating (16.100) over the region defined by

$$0 < u_1 < u_2 < v_1 < v_2 < v_3 < \dots < 1$$

i.e., integrating  $u_1$  over 0 to  $u_2$ ; then  $u_2$  over 0 to  $v_1$  and so on. The value of the integral will, on simplification, come out to be

$$\frac{n_1! n_2!}{(n_1 + n_2)!} = \frac{1}{\binom{n_1 + n_2}{n_1}}$$

Since there are exactly  $\binom{n_1 + n_2}{n_1}$  arrangements of  $n_1$ , x's and  $n_2$ , y's, it follows that all the arrangements of x's and y's are equally likely.

Since under  $H_0$  all the  $\binom{n_1 + n_2}{n_1}$  arrangements of  $n_1$  x's and  $n_2$  y's are equally likely, to obtain the distribution of  $U$  under  $H_0$ , it is necessary to count all the arrangements with exactly 'u' runs. Let us first take the case of even number of runs, i.e.,  $u = 2k$ . In this case we should have  $k$  runs of x's and  $k$  runs of y's.

$n_1$  x's will give  $k$  runs if they are separated by  $(k - 1)$  vertical bars in distinct spaces between the x's. In other words,  $(k - 1)$  spaces are to come out of the total number of  $(n_1 - 1)$  spaces between the  $n_1$  x's and this can happen in  $\binom{n_1 - 1}{k - 1}$  ways. Hence  $k$  runs of x's can be obtained in  $\binom{n_1 - 1}{k - 1}$  ways.

Similarly,  $k$  runs of y's can be obtained in  $\binom{n_2 - 1}{k - 1}$  ways.

The same result holds if the sequence of runs in (16.99) starts with x or with y. Since a sequence of type (16.99) may start with x or y, we get

$$P(U = 2k) = \frac{2 \left( \binom{n_1 - 1}{k - 1} \right) \left( \binom{n_2 - 1}{k - 1} \right)}{\binom{n_1 + n_2}{n_1}}$$

If the number of runs in (16.99) is odd, i.e.,  $u = 2k + 1$ , then we should have either (i)  $(k + 1)$  runs of x and  $k$  runs of y or (ii)  $k$  runs of x and  $(k + 1)$  runs of y. Hence

$$P(U = 2k + 1) = P(i) + P(ii)$$

$$= \frac{\left( \binom{n_1 - 1}{k} \right) \left( \binom{n_2 - 1}{k - 1} \right) + \left( \binom{n_1 - 1}{k - 1} \right) \left( \binom{n_2 - 1}{k} \right)}{\binom{n_1 + n_2}{n_1}}$$

Hence the distribution of  $U$  under  $H_0$  is given by

$$\left. \begin{aligned} P(U=2k) &= \frac{2 \binom{n_1-1}{k-1} \binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}, \text{ and} \\ P(U=2k+1) &= \frac{\binom{n_1-1}{k} \binom{n_2-1}{k-1} + \binom{n_1-1}{k-1} \binom{n_2-1}{k}}{\binom{n_1+n_2}{n_1}} \end{aligned} \right\} \quad \dots(16-101)$$

If the probability of type I error is fixed as  $\alpha$ , then  $u_0$  is determined from the equation :

$$\sum_{u=2}^{u_0} h(u) = \alpha \quad \dots(16-102)$$

where  $h(u)$  is the probability function of  $U$  given by (16-101).

Calculation of  $u_0$  from (16-102) is quite tedious and cumbersome unless  $n_1$  and  $n_2$  are large in which case under  $H_0$ ,  $U$  is asymptotically normal with

$$E(U) = \frac{2n_1 n_2}{n_1 + n_2} + 1 \quad \dots(16-103)$$

$$\text{Var}(U) = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} \quad \dots(16-104)$$

and we can use the normal test

$$Z = \frac{U - E(U)}{\sqrt{\text{Var}(U)}} \sim N(0, 1), \text{ asymptotically.} \quad \dots(16-105)$$

This approximation is fairly good if each of  $n_1$  and  $n_2$  is greater than 10. Since the alternative hypothesis is "too few runs", the test is ordinarily one-tailed with only negative values leading to the rejection of  $H_0$ .

**16-8-4. Test for Randomness.** Another application of the 'run' theory is in testing the randomness of a given set of observations. Let  $x_1, x_2, \dots, x_n$  be the set of observations arranged in the order in which they occur, i.e.,  $x_i$  is the  $i$ th observation in the outcome of an experiment. Then, for each of the observations, we see if it is above or below the value of the median of the observations and write  $A$  if the observation is above and  $B$  if it is below the median value. Thus we get a sequence of  $A$ 's and  $B$ 's of the type, (say),

$$A \ B \ B \ A \ A \ A \ B \ A \ B \ B \quad \dots(*)$$

Under the null hypothesis  $H_0$  that the set of observations is random, the number of runs  $U$  in (\*) is a r.v. with

$$E(U) = \frac{n+2}{2} \quad \text{and} \quad \text{Var}(U) = \frac{n}{4} \left( \frac{n-2}{n-1} \right) \quad \dots(16-106)$$

For large  $n$  (say,  $> 25$ ),  $U$  may be regarded as asymptotically normal and we may use the normal test.

**16-8-5. Median Test.** Median test is a statistical procedure for testing if two independent ordered samples differ in their central tendencies. In other words, it gives information if two independent samples are likely to have been drawn from the populations with the same median.

As, in 'run' test, let  $x_1, x_2, \dots, x_{n_1}$  and  $y_1, y_2, \dots, y_{n_2}$  be two independent ordered samples from the populations with p.d.f.'s  $f_1(\cdot)$  and  $f_2(\cdot)$  respectively. The measurements must be at least ordinal. Let  $z_1, z_2, \dots, z_{n_1+n_2}$  be the combined ordered sample. Let  $m_1$  be the number of  $x$ 's and  $m_2$  the number of  $y$ 's exceeding the median value  $M$ , (say), of the combined sample.

Then under the null hypothesis that the samples come from the same population or from different populations with the same median, i.e., under  $H_0 : f_1(\cdot) = f_2(\cdot)$ , the joint distribution of  $m_1$  and  $m_2$  is the hypergeometric distribution with probability function :

$$p(m_1, m_2) = \frac{\binom{n_1}{m_1} \binom{n_2}{m_2}}{\binom{n_1 + n_2}{m_1 + m_2}} \quad \dots(16-107)$$

If  $m_1 < n_1/2$ , then the critical region corresponding to the size of type 1 error  $\alpha$ , is given by  $m_1 < m_1'$  where  $m_1'$  is computed from the equation

$$\sum_{m_1=1}^{m_1'} p(m_1, m_2) = \alpha \quad \dots(16-108)$$

The distribution of  $m_1$  under  $H_0$  is also hyper-geometric with

$$\left. \begin{aligned} E(m_1) &= \frac{n_1}{2}, \text{ if } N = n_1 + n_2 \text{ is even} \\ &= \frac{n_1}{2} \cdot \frac{N-1}{N}, \text{ if } N \text{ is odd} \\ \text{and } \text{Var}(m_1) &= \frac{n_1 n_2}{4(N-1)}, \text{ if } N \text{ is even} \\ &= \frac{n_1 n_2 (N+1)}{4N^2}, \text{ if } N \text{ is odd} \end{aligned} \right\} \quad \dots(16-109)$$

This distribution is most of the times quite inconvenient to use. However for large samples, we may regard  $m_1$  to be asymptotically normal and use normal test, viz.,

$$Z = \frac{m_1 - E(m_1)}{\sqrt{\text{Var}(m_1)}} \sim N(0, 1), \text{ asymptotically.} \quad \dots(16-110)$$

**Remarks 1.** The observations  $m_1$  and  $m_2$  can be classified into the following  $2 \times 2$  contingency table.

	<i>Sample 1</i>	<i>Sample 2</i>	<i>Total</i>
No. of observations $> M$	$m_1$	$m_2$	$m_1 + m_2$
No. of observations $< M$	$n_1 - m_1$	$n_2 - m_2$	$n_1 + n_2 - m_1 - m_2$
Total	$n_1$	$n_2$	$n_1 + n_2 = N$

If frequencies are small we can compute the exact probabilities from (16-107), rather than approximate them. However, if frequencies are large we may use  $\chi^2$ -test with 1 d.f. (for a  $2 \times 2$  contingency table) for testing  $H_0$ .

The approximation is fairly good if both  $n_1$  and  $n_2$  exceed 10.

2. Median test is sensitive to the differences in location between  $f_1(x)$  and  $f_2(y)$  but not to differences in their shapes. Thus if  $f_1(x)$  and  $f_2(y)$  have the same median, we would expect  $H_0 : f_1(\cdot) = f_2(\cdot)$  to be accepted ordinarily even though their shapes are quite different.

3. Generally, the median test makes the correct decision with a little more assurance than does the sign test (c.f. §. 16-8-6) but not as decisively as the  $t$ -test.

**16-8-6. Sign Test.** Consider a situation where it is desired to compare two things or materials under various sets of conditions. An experiment is thus conducted under the following circumstances :

(i) When there are pairs of observations on two things being compared.

(ii) For any given pair, each of the two observations is made under similar extraneous conditions.

(iii) Different pairs are observed under different conditions.

Condition (iii) implies that the differences  $d_i = x_i - y_i ; i = 1, 2, \dots, n$  have different variances and thus renders the paired  $t$ -test (Chapter 14) invalid, which would have otherwise been used unless there was obvious non-normality. So, in such a case we use the 'Sign Test', named so since it is based on the signs (plus or minus) of the deviations  $d_i = x_i - y_i$ . No assumptions are made regarding the parent population. The only assumptions are :

(i) Measurements are such that the deviations  $d_i = x_i - y_i$ , can be expressed in terms of positive or negative signs.

(ii) Variables have continuous distribution.

(iii)  $d_i$ 's are independent.

Different pairs  $(x_i, y_i)$  may be from different populations (say w.r.t. age, weight, stature, education, etc.). The only requirement is that within each pair, there is matching w.r.t relevant extraneous factors.

**Procedure.** Let  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  be  $n$  paired sample observations drawn from the two populations with p.d.f.'s  $f_1(\cdot)$  and  $f_2(\cdot)$ . We want to test the null hypothesis  $H_0 : f_1(\cdot) = f_2(\cdot)$ . To test  $H_0$ , consider  $d_i = x_i - y_i$ , ( $i = 1, 2, \dots, n$ ). When  $H_0$  is true,  $x_i$  and  $y_i$  constitute a random sample of size 2 from the same population. Since the probability that the first of the two sample observations exceeds the second is same as the probability that the second exceeds the first and since hypothetically the probability of a tie is zero,  $H_0$  may be restated as :

$$H_0: P[X - Y > 0] = \frac{1}{2} \text{ and } P[X - Y < 0] = \frac{1}{2}$$

Let us define:

$$U_i = \begin{cases} 1, & \text{if } x_i - y_i > 0 \\ 0, & \text{if } x_i - y_i < 0 \end{cases}$$

$U_i$  is a Bernoulli variate with  $p = P(x_i - y_i > 0) = \frac{1}{2}$ . Since  $U_i$ 's,  $i = 1, 2, \dots, n$  are independent,  $U = \sum_{i=1}^n U_i$ , the total number of positive deviations, is a Binomial variate with parameters  $n$  and  $p (= \frac{1}{2})$ . Let the number of positive deviations be  $k$ . Then

$$\begin{aligned} P(U \leq k) &= \sum_{r=0}^k \binom{n}{r} p^r q^{n-r}, \quad (p = q = \frac{1}{2} \text{ under } H_0). \\ &= \left( \frac{1}{2} \right)^n \sum_{r=0}^k \binom{n}{r} = p', \text{ (say).} \end{aligned} \quad \dots(16-111)$$

If  $p' \leq 0.05$ , we reject  $H_0$  at 5% level of significance and if  $p' > 0.05$ , we conclude that the data do not provide any evidence against the null hypothesis, which may therefore, be accepted.

For large samples, ( $n \geq 30$ ), we may regard  $U$  to be asymptotically normal with, (under  $H_0$ )

$$\therefore E(U) = np = n/2 \text{ and } \text{Var}(U) = npq = n/4.$$

$$\therefore Z = \frac{U - E(U)}{\sqrt{\text{Var}(U)}} = \frac{U - n/2}{\sqrt{(n/4)}} \text{ is asymptotically } N(0, 1). \quad \dots(16-112)$$

and we may use normal test.

**16-8-7. Mann-Whitney-Wilcoxon U-test.** This non-parametric test for two samples was described by Wilcoxon and studied by Mann and Whitney. It is the most widely used test as an alternative to the  $t$ -test when we do not make the  $t$ -test assumptions about the parent population.

Let  $x_i$  ( $i = 1, 2, \dots, n_1$ ) and  $y_j$  ( $j = 1, 2, \dots, n_2$ ) be independent ordered samples of size  $n_1$  and  $n_2$  from the populations with p.d.f.  $f_1(\cdot)$  and  $f_2(\cdot)$  respectively. We want to test the null hypothesis  $H_1 : f_1(\cdot) = f_2(\cdot)$ . Like the run test, Mann-Whitney test is based on the pattern of the  $x$ 's and  $y$ 's in the combined ordered sample. Let  $T$  denote the sum of ranks of the  $y$ 's in the

*combined ordered sample.* For example, for the pattern (16-99) on page 16-61 of combined ordered sample the ranks of  $y$  observations are respectively 3, 4, 5, 7, 10 etc. and  $T = 3 + 4 + 5 + 7 + 10 + \dots$  The test statistic  $U$  is then defined in terms of  $T$  as follows :

$$U = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - T \quad \dots(16-113)$$

If  $T$  is significantly large or small then  $H_0 : f_1(\cdot) = f_2(\cdot)$  is rejected. The problem is to find the distribution of  $T$  under  $H_0$ . Unfortunately, it is very troublesome to obtain the distribution of  $T$  under  $H_0$ . However, Mann and Whitney have obtained the distribution of  $T$  for small  $n_1$  and  $n_2$ , have found the moments of  $T$  in general and shown that  $T$  is asymptotically normal. It has been established that under  $H_0$ ,  $U$  is asymptotically normally distributed as  $N(\mu, \sigma^2)$ , where

$$\begin{aligned}\mu &= E(U) = \frac{n_1 n_2}{2} \\ \sigma^2 &= \text{Var}(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \quad \dots(16-114)\end{aligned}$$

Hence

$$Z = \frac{U - \mu}{\sigma} \sim N(0, 1), \text{ asymptotically.} \quad \dots(16-114a)$$

and normal test can be used. The approximation is fairly good if both  $n_1$  and  $n_2$  are greater than 8.

**Remark.** The asymptotic relative efficiency (ARE) of Mann-Whitney's  $U$ -test relative to two samples  $t$ -test is greater than or equal to 0.864. For a normal population, this  $ARE = 3/\pi = 0.955$ . Accordingly, Mann-Whitney's  $U$ -test is regarded as the best non-parametric test for location.

### EXERCISE 16 (c)

1. Explain what is meant by non-parametric methods. How do they differ from parametric methods ? Illustrate your answer by considering a suitable nonparametric test for the hypothesis that two independent samples have come from the same population.
2. (a) Derive the sign test, stating clearly the assumptions made.  
(b) Describe the median test for the two-sample location problem. Find the distribution of the test statistic and compute its mean and variance under the null hypothesis. How is the test carried out in case of large samples ?
3. Explain the main difference between parametric and non-parametric approaches to the theory of statistical inference. Derive the sign test for two sample problem. *(Delhi Univ. B.Sc. (Stat. Hons.), 1988)*
4. Describe the sign test.

$X_1, X_2, \dots, X_{10}$  is a random sample of size 10 from a population having distribution function  $F(x)$ . Test the hypothesis  $H_0 : F(72) = \frac{1}{2}$  against the alternative hypothesis,  $H_1 : F(72) > \frac{1}{2}$ . *[Madras Univ. B.Sc., 1988]*

**5. Explain Median Test and how it is applied.**

The observations of a random sample of size 10 from a distribution which is symmetric about  $K_5$  are 20.2, 24.1, 21.3, 17.2, 19.8, 16.5, 21.8, 18.7, 17.1, 19.9. Use Wilcoxon's Test to test the hypothesis  $H_0 : K_5 = 18$  against  $H_1 : K_5 > 18$  if  $\alpha = 0.05$ . You may use the normal approximation.

[*Agra Univ. B.Sc., 1989*]

**6. Describe the procedure in median test when there are two independent samples. What non-parametric test would you use when the two samples are related.**

**7. Discuss the Mann-Whitney-Wilcoxon test for the equality of two population distribution functions.** [*Delhi Univ. B.Sc. (Stat. Hons.), 1986*]

**8. What are the advantages and disadvantages of non-parametric methods over parametric methods ?**

Develop the following non-parametric tests, stating the underlying assumptions and the null hypotheses :

(a) Median test

(b) Mann-Whitney-Wilcoxon test.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1993*]

**9. Explain the main difference between the parametric and non-parametric approaches to the theory of statistical inference. What are the advantages of non-parametric tests ? Develop Median test and Mann-Whitney-Wilcoxon test.**

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1985*]

**10. Distinguish between 'sign test' and 'Wilcoxon signed rank test'. Describe the sign test for testing that the population median is  $M_0$  against the alternative that the median is  $M_1 (> M_0)$ .**

**11. Develop the Mann-Whitney-Wilcoxon test and obtain the mean and variance of the test statistic  $T$ . How is the test carried out for large samples ?**

**12. Explaining the distinction between the parametric and non-parametric tests, write down the advantages of non-parametric tests. Also write their disadvantages.**

Thirty observations as given below are obtained :

24, 35, 12, 50, 60, 70, 68, 49, 80, 25, 69, 28, 28, 11, 83,

31, 37, 34, 54, 75, 45, 95, 75, 26, 43, 57, 94, 48, 63, 45

Test their randomness by considering the sequence of positive and negative signs. [*Agra Univ. B.Sc., 1989*]

**13. What are the advantages and disadvantages of Non-Parametric Methods over Parametric Methods ?**

Derive the Wald-Wolfowitz run test for testing the equality of two distribution functions. [*Delhi Univ. B.Sc. (Stat. Hons.), 1987*]

**14. What are the advantages of Non Parametric tests ? Define a run and the length of a run. Describe the Run Test in detail for testing the equality of the two populations and extend the test when the ties occur.**

[*Delhi Univ. B.Sc. (Stat. Hons.), 1983*]

**15. What are runs ? Comment on their utility in non-parametric inference.**

If  $R_1$  and  $R_2$  denote the number of runs of  $n_1$  objects of one type and  $n_2$  objects of another type in a sample of size  $n_1 + n_2$ , then find the probability that  $R_1 + R_2 = r$ , for  $r$  even and  $r$  odd, and also the mean and variance of  $R_1 + R_2$  when all these  $n_1 + n_2$  observations arise from the same distribution.

[*Delhi Univ. M.Sc. (Stat.), 1991*]

16. (i) Explain how the run test can be used to test randomness.

(ii) In the median test with samples of size 9 and 7 respectively, from two populations find the probability density function of the random variable representing the number of values of the samples from the first population in the lower half of the combined sample. [*Madras Univ. B.Sc., 1988*]

17. (a) The win-loss record of a certain basketball team for its last 50 consecutive games was as follows :—

W W W W W W L W W W W W W L W L W W W L L W W W W  
L W W W L L W W W W W W L L W W L L L W W L W W W W

Apply run test to test that sequence of wins and losses is random.

(b) Use an appropriate non-parametric test procedure to test for randomness the following set of 30 two-digit numbers :

15,	17,	01,	65,	69,	69,	58	41,	81,	16,
16,	20,	00,	84,	22,	28,	26,	46,	66,	36,
86,	66,	17,	34,	49,	85,	45,	51,	40,	10.

18. At the beginning of the year a first grade class was randomly divided into two groups. One group was taught to read using a uniform method, where all students progressed from one stage to the next at the same time, following the teacher's direction. The second group was taught to read using an individual method, where each student progressed at his own rate according to a programmed work book, under supervision of the teacher. At the end of the year each student was given a reading ability test with the following results :

<i>First Group</i>			<i>Second Group</i>		
227	55	184	202	271	63
176	234	147	14	151	284
252	194	88	165	235	53
149	247	161	171	147	228
16	92	171	292	99	271

Use the Wald-Wolfowitz run-test to test for the equality of the distribution functions of the two groups:

19. Using the number of runs above and below the median, test for randomness the following set of a table of 2-digit numbers :

15, 77, 01, 65, 69, 69, 58, 40, 81, 16, 16 20, 00, 84, 22,  
28, 26, 46, 66, 36, 86, 66, 17, 43, 49, 85, 40, 51, 40, 10.

16.9. Sequential Analysis – Introduction We have seen that in Neyman-Pearson theory of testing hypothesis,  $n$ , the sample size is regarded as a fixed constant and keeping  $\alpha$  fixed, we minimise  $\beta$ . But in the sequential analysis theory propounded by A. Wald  $n$ , the sample size is not fixed but is regarded as a random variable whereas both  $\alpha$  and  $\beta$  are fixed constants.

16.9.1. Sequential Probability Ratio Test (SPRT). The best known procedure in sequential testing is the *Sequential Probability Ratio Test* (SPRT) developed by A. Wald discussed below.

Suppose we want to test the hypothesis,  $H_0 : \theta = \theta_0$  against the alternative hypothesis,  $H_1 : \theta = \theta_1$ , for a distribution with p.d.f.  $f(x, \theta)$ . For any positive integer  $m$ , the likelihood function of a sample  $x_1, x_2, \dots, x_m$  from the population with p.d.f.  $f(x, \theta)$  is given by

$$L_{1m} = \prod_{i=1}^m f(x_i, \theta_1) \text{ when } H_1 \text{ is true,}$$

and by  $L_{0m} = \prod_{i=1}^m f(x_i, \theta_0)$  when  $H_0$  is true,

and the likelihood ratio  $\lambda_m$  is given by

$$\lambda_m = \frac{L_{1m}}{L_{0m}} = \frac{\prod_{i=1}^m f(x_i, \theta_1)}{\prod_{i=1}^m f(x_i, \theta_0)}, (m = 1, 2, \dots) \quad \dots(16-115)$$

The SPRT for testing  $H_0$  against  $H_1$  is defined as follows :

At each stage of the experiment (at the  $m$ th trial for any integral value  $m$ ), the likelihood ratio  $\lambda_m$ , ( $m = 1, 2, \dots$ ) is computed.

- (i) If  $\lambda_m \geq A$ , we terminate the process with the rejection of  $H_0$
  - (ii) If  $\lambda_m \leq B$ , we terminate the process with the acceptance of  $H_0$ , and
  - (iii) If  $B < \lambda_m < A$ , we continue sampling by taking an additional observation.
- ... (16-116)

Here  $A$  and  $B$ , ( $B < A$ ) are the constants which are determined by the relation

$$A = \frac{1 - \beta}{\alpha}, B = \frac{\beta}{1 - \alpha} \quad \dots(16-117)$$

where  $\alpha$  and  $\beta$  are the probabilities of type I error and type II error respectively.

From computational point of view, it is much convenient to deal with  $\log \lambda_m$  rather than  $\lambda_m$ , since

$$\log \lambda_m = \sum_{i=1}^m \log \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)} = \sum_{i=1}^m z_i \quad \dots(16-118)$$

$$\text{where } z_i = \log \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)} \quad \dots(16-118a)$$

In terms of  $z_i$ 's, SPRT is defined as follows :

- (i) If  $\sum z_i \geq \log A$ , reject  $H_0$
  - (ii) If  $\sum z_i \leq \log B$ , reject  $H_1$
  - (iii) If  $\log B < \sum z_i < \log A$ , continue sampling by taking an additional observation.
- ... (16-119)

**Remarks 1.** In SPRT, we continue taking additional observations unless the inequality

$$B < \lambda_m < A \Rightarrow \log B < \sum z_i < \log A,$$

is violated at either end. It has been proved that SPRT eventually terminates with probability one.

2. Sequential schemes provide for a minimum amount of sampling and thus result in considerable saving in terms of inspection, time and money. As compared with single sampling, sequential scheme requires on the average 33% to 50% less inspection for the same degree of protection i.e., for the same values of  $\alpha$  and  $\beta$ .

**16-9-2. Operating Characteristic (O.C.) Function of SPRT.** The O.C. function  $L(\theta)$  is defined as

$L(\theta)$  = Probability of accepting  $H_0 : \theta = \theta_0$  when  $\theta$  is the true value of the parameter,

and since the power function

$P(\theta)$  = Probability of rejecting  $H_0$  where  $\theta$  is the true value, we get

$$L(\theta) = 1 - P(\theta) \quad \dots(16-120)$$

The O.C. function of a SPRT for testing  $H_0 : \theta = \theta_0$  against the alternative  $H_1 : \theta = \theta_1$ , in sampling from a population with density function  $f(x, \theta)$  is given by

$$L(\theta) = \frac{A^{h(\theta)} - 1}{A^{h(\theta)} - B^{h(\theta)}} \quad \dots(16-121)$$

where for each value of  $\theta$ , the value of  $h(\theta) \neq 0$ , is to be determined so that

$$E\left[\frac{f(x, \theta_1)}{f(x, \theta_0)}\right]^{h(\theta)} = 1 \quad \dots(16-122)$$

where the constants  $A$  and  $B$  have already been defined in (16-117). It has been proved that under very simple conditions on the nature of the function  $f(x, \theta)$ , there exists a unique value of  $h(\theta) \neq 0$  such that (16-122) is satisfied.

**16-9-3. Average Sample Number (A.S.N.).** The sample size  $n$  in sequential testing is a random variable which can be determined in terms of the true density function  $f(x, \theta)$ . The A.S.N. function for the S.P.R.T. for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , is given by

$$E(n) = \frac{L(\theta) \log B + [1 - L(\theta)] \log A}{E(Z)} \quad \dots(16-123)$$

$$\text{where } Z = \log \left[ \frac{f(x, \theta_1)}{f(x, \theta_0)} \right], A = \frac{1 - \beta}{\alpha}, B = \frac{\beta}{1 - \alpha} \quad \dots(16-123a)$$

**Example 16-11.** Give the S.P.R.T. for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1 (> \theta_0)$ , in sampling from a normal density.

$$\frac{x}{\theta} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2\right], -\infty < x < \infty$$

where  $\sigma$  is known. Also obtain its O.C. function and A.S.N. function.

$$\text{Solution. } \frac{f(x_i; \theta_1)}{f(x_i; \theta_0)} = \exp \left[ -\frac{1}{2\sigma^2} \{ (x_i - \theta_1)^2 - (x_i - \theta_0)^2 \} \right] \\ = \exp \left[ -\frac{1}{2\sigma^2} \{ (\theta_0 - \theta_1) (2x_i - \theta_0 - \theta_1) \} \right] \quad \dots (*)$$

$$\therefore z_i = \log \frac{f(x_i; \theta_1)}{f(x_i; \theta_0)} = \frac{\theta_1 - \theta_0}{\sigma^2} \left[ x_i - \frac{\theta_0 + \theta_1}{2} \right] \quad \dots (**)$$

$$\Rightarrow \log \lambda_m = \sum_{i=1}^m z_i = \frac{\theta_1 - \theta_0}{\sigma^2} \left[ \sum_i x_i - \frac{m(\theta_0 + \theta_1)}{2} \right]$$

Hence the S.P.R.T. for  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , is given by [c.f. (16-119)] :

(i) Reject  $H_0$  if

$$\frac{\theta_1 - \theta_0}{\sigma^2} \left[ \sum x_i - \frac{m(\theta_0 + \theta_1)}{2} \right] \geq \log \left( \frac{1-\beta}{\alpha} \right) \\ \Rightarrow \sum_{i=1}^m x_i \geq \frac{\sigma^2}{\theta_1 - \theta_0} \log \left( \frac{1-\beta}{\alpha} \right) + \frac{m(\theta_0 + \theta_1)}{2}; (\theta_1 > \theta_0)$$

(ii) Accept  $H_0$  if

$$\frac{\theta_1 - \theta_0}{\sigma^2} \left[ \sum x_i - \frac{m(\theta_0 + \theta_1)}{2} \right] \leq \log \left( \frac{\beta}{1-\alpha} \right) \\ \Rightarrow \sum_{i=1}^m x_i \leq \frac{\sigma^2}{\theta_1 - \theta_0} \log \left( \frac{\beta}{1-\alpha} \right) + \frac{m(\theta_0 + \theta_1)}{2}; (\theta_1 > \theta_0)$$

and (iii) Continue taking additional observations as long as

$$\log \left( \frac{\beta}{1-\alpha} \right) < \frac{\theta_1 - \theta_0}{\sigma^2} \left[ \sum x_i - \frac{m(\theta_0 + \theta_1)}{2} \right] < \log \left( \frac{1-\beta}{\alpha} \right) \\ \Rightarrow \frac{\sigma^2}{\theta_1 - \theta_0} \log \left( \frac{1-\beta}{\alpha} \right) + \frac{m(\theta_0 + \theta_1)}{2} < \sum x_i < \frac{\sigma^2}{\theta_1 - \theta_0} \log \left( \frac{\beta}{1-\alpha} \right) \\ + \frac{m(\theta_0 + \theta_1)}{2}$$

**O.C. Function.** First of all we shall determine  $h = h(\theta) \neq 0$ , from (16-122) i.e., from

$$\int_{-\infty}^{\infty} \left[ \frac{f(x; \theta_1)}{f(x; \theta_0)} \right]^h f(x; \theta) dx = 1$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2\right] \cdot \left[ \exp\left\{-\frac{1}{2\sigma^2}(\theta_1 - \theta_0) \times (-2x + \theta_0 + \theta_1)\right\} \right]^h dx = 1, [\text{On using (*)}]$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\{x^2 - 2x((\theta_1 - \theta_0)h + \theta) + \theta^2 + (\theta_1^2 - \theta_0^2)h\}\right] dx = 1$$

If we take

$$\begin{aligned} \lambda &= (\theta_1 - \theta_0)h + \theta \\ \lambda^2 &= (\theta_1^2 - \theta_0^2)h + \theta^2 \end{aligned} \quad \dots (***)$$

then L.H.S. becomes

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x - \lambda)^2\right] dx,$$

which being the total area under normal probability curve with mean  $\lambda$  and variance  $\sigma^2$  is always unity, as desired. Thus  $h = h(\theta)$  is the solution of (\*\*\*), and is given by

$$\begin{aligned} (\theta_1^2 - \theta_0^2)h + \theta^2 &= [\theta_1 - \theta_0]h + \theta]^2 \\ \Rightarrow (\theta_1^2 - \theta_0^2)h &= (\theta_1 - \theta_0)^2h^2 + 2\theta(\theta_1 - \theta_0)h \end{aligned}$$

Since  $h = h(\theta) \neq 0$  and  $\theta_1 \neq \theta_0$ , on dividing throughout by  $(\theta_1 - \theta_0)h$ , we get

$$\begin{aligned} (\theta_1 + \theta_0) &= (\theta_1 - \theta_0)h + 2\theta \\ \Rightarrow h(\theta) &= \frac{\theta_1 + \theta_0 - 2\theta}{\theta_1 - \theta_0} \end{aligned}$$

Substituting for  $h(\theta)$  in (16.121) we get the required expression for the O.C. function.

**A.S.N. function.** We have

$$Z = \log \frac{f(x, \theta_1)}{f(x, \theta_0)} = \frac{\theta_1 - \theta_0}{\sigma^2} \left[ x - \frac{\theta_0 + \theta_1}{2} \right] \quad [\text{From (**)}]$$

$$\begin{aligned} \therefore E(Z) &= \frac{\theta_1 - \theta_0}{2\sigma^2} [2E(x) - \theta_0 - \theta_1] \\ &= \frac{\theta_1 - \theta_0}{2\sigma^2} [2\theta - \theta_0 - \theta_1] \end{aligned}$$

Substituting in (16.123), we get the required A.S.N. function.

**Example 16.12.** Let  $X$  have the distribution :

$$f(x, \theta) = \theta^x (1-\theta)^{1-x}; x = 0, 1; 0 < \theta < 1$$

For testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , construct S.P.R.T. and obtain its A.S.N. and O.C. functions.

**Solution.** We have

$$\begin{aligned}\lambda_m &= \frac{L(x_1, x_2, \dots, x_m | H_1)}{L(x_1, x_2, \dots, x_m | H_0)} \\ &= \left\{ \theta_1^{\sum_{i=1}^m x_i} (1 - \theta_1)^{m - \sum_{i=1}^m x_i} \right\} + (\theta_0^{\sum x_i} (1 - \theta_0)^{m - \sum x_i}) \\ &= \left( \frac{\theta_1}{\theta_0} \right)^{\sum_{i=1}^m x_i} \left( \frac{1 - \theta_1}{1 - \theta_0} \right)^{m - \sum_{i=1}^m x_i}.\end{aligned}$$

$$\begin{aligned}\log \lambda_m &= \sum x_i \log (\theta_1/\theta_0) + (m - \sum x_i) \log \left( \frac{1 - \theta_1}{1 - \theta_0} \right) \\ &= \sum_{i=1}^m x_i \log \left[ \frac{\theta_1 (1 - \theta_0)}{\theta_0 (1 - \theta_1)} \right] + m \log \left( \frac{1 - \theta_1}{1 - \theta_0} \right)\end{aligned}$$

Hence SPRT for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , is given by [c.f. (16-119)] :

(i) Accept  $H_0$  if  $\log \lambda_m \leq \log \left( \frac{\beta}{1 - \alpha} \right) = b$ , (say)

$$\text{i.e., if } \sum_{i=1}^m x_i \leq \frac{b - m \log [(1 - \theta_1)/(1 - \theta_0)]}{\log [\theta_1(1 - \theta_0)/\theta_0(1 - \theta_1)]} = a_m, \text{ (say).}$$

(ii) Reject  $H_0$  (Accept  $H_1$ ) if  $\log \lambda_m \geq \log \frac{1 - \beta}{\alpha} = r_m$ , (say)

$$\text{i.e., if } \sum_{i=1}^m x_i \geq \frac{a_m - m \log [(1 - \theta_1)/(1 - \theta_0)]}{\log [\theta_1(1 - \theta_0)/\theta_0(1 - \theta_1)]} = r_m, \text{ (say).}$$

(iii) Continue sampling if

$$b < \log \lambda_m < a \Rightarrow a_m < \sum x_i < r_m$$

**O.C. Function.** O.C. function is given by :

$$L(\theta) = [A^{h(\theta)} - 1] / [A^{h(\theta)} - B^{h(\theta)}] \quad [\text{c.f. (16-121)}] \dots(i)$$

where for each value of  $\theta$ ,  $h(\theta) \neq 0$  is to be determined such that

$$\begin{aligned}&E \left[ \frac{f(x, \theta_1)}{f(x, \theta_0)} \right]^{h(\theta)} = 1 \quad [\text{c.f. (16-122)}] \\ \Rightarrow &\sum_{x=0}^1 \left[ \frac{f(x, \theta_1)}{f(x, \theta_0)} \right]^{h(\theta)} f(x, \theta) = 1 \\ \Rightarrow &\sum_{x=0}^1 \left[ \left( \frac{\theta_1}{\theta_0} \right)^x \left( \frac{1 - \theta_1}{1 - \theta_0} \right)^{1-x} \right]^{h(\theta)} \theta^x (1 - \theta)^{1-x} = 1 \\ \Rightarrow &\left( \frac{1 - \theta_1}{1 - \theta_0} \right)^{h(\theta)} \cdot (1 - \theta) + \left( \frac{\theta_1}{\theta_0} \right)^{h(\theta)} \cdot \theta = 1 \quad \dots(ii)\end{aligned}$$

The solution of this equation for  $h = h(\theta)$  is very tedious. From practical point of view, instead of solving (ii) for  $h$  we regard  $h$  as a parameter and solve it for  $\theta$ , thus giving

$$\theta \left[ \left( \frac{\theta_1}{\theta_0} \right)^{h(\theta)} - \left( \frac{1 - \theta_1}{1 - \theta_0} \right)^{h(\theta)} \right] = 1 - \left( \frac{1 - \theta_1}{1 - \theta_0} \right)^{h(\theta)}$$

$$\Rightarrow \theta = \frac{1 - [(1 - \theta_1)/(1 - \theta_0)]^{h(\theta)}}{(\theta_1/\theta_0)^{h(\theta)} - [(1 - \theta_1)/(1 - \theta_0)]^{h(\theta)}} = \theta(h), \text{ (say). } \dots(iii)$$

Using (i), we have

$$L(\theta) = \frac{[(1 - \beta)/\alpha]^h - 1}{[(1 - \beta)/\alpha]^h - [\beta/(1 - \alpha)]^h} = L(\theta, h), \text{ (say). } \dots(iv)$$

Various points on the O.C. curve are obtained by assigning arbitrary values to ' $h$ ' and computing the corresponding values of  $\theta$  and  $L(\theta)$  from (iii) and (iv) respectively.

### A.S.N. Function.

$$Z = \log \left[ \frac{f(x, \theta_1)}{f(x, \theta_0)} \right]; A = \frac{1 - \beta}{\alpha}, B = \frac{\beta}{1 - \alpha}$$

$$\therefore E(Z) = \sum_{x=0}^1 \log \left[ \frac{f(x, \theta_1)}{f(x, \theta_0)} \right] \cdot f(x, \theta)$$

$$= \sum_{x=0}^1 \log \left[ \left( \frac{\theta_1}{\theta_0} \right)^x \left( \frac{1 - \theta_1}{1 - \theta_0} \right)^{1-x} \right] \cdot \theta^x (1 - \theta)^{1-x}$$

$$= (1 - \theta) \log \left( \frac{1 - \theta_1}{1 - \theta_0} \right) + \theta \cdot \log \left( \frac{\theta_1}{\theta_0} \right)$$

$$= \theta \log \left[ \frac{\theta_1 (1 - \theta_0)}{\theta_0 (1 - \theta_1)} \right] + \log \left( \frac{1 - \theta_1}{1 - \theta_0} \right) \quad \dots(v)$$

A.S.N. is given by

$$E(n) = \frac{L(\theta) \log B + [1 - L(\theta)] \cdot \log A}{E(Z)} \quad \dots(vi)$$

Substituting the values of  $E(Z)$  and  $L(\theta)$  from (v) and (iv) in (vi), we get the A.S.N. function.

**Remark.** If  $h$  assumes negative values i.e., if instead of  $h$  we take  $-h$  where  $h > 0$ , then

$$L(\theta, -h) = \frac{A^{-h} - 1}{A^{-h} - B^{-h}} = \left( \frac{1 - A^h}{B^h - A^h} \right) B^h = \left( \frac{A^h - 1}{A^h - B^h} \right) \cdot B^h$$

$$\Rightarrow L(\theta, -h) = B^h \cdot L(\theta, h) \quad \dots(vii)$$

and

$$\theta(-h) = \frac{[(1 - \theta_1)/(1 - \theta_0)]^h - 1}{[(1 - \theta_1)/(1 - \theta_0)]^h - (\theta_1/\theta_0)^h} \left( \frac{\theta_1}{\theta_0} \right)^h$$

$$= \theta(h) \cdot \left( \frac{\theta_1}{\theta_0} \right)^h \quad \dots(viii)$$

Formulae (vii) and (viii) are very convenient to use for obtaining the points on O.C. curve for arbitrary negative values of  $h$ .

### EXERCISE 16(d)

1. (a) Describe Wald's Sequential Probability Ratio Test.
- (b) Explain how the sequential test procedure differs from the Neyman-Pearson test procedure.
2. Define the *OC* function and *ASN* function in sequential analysis. Derive their approximate expressions for the sequential probability ratio test of a simple hypothesis against a simple alternative.

3. Describe Wald's S.P.R.T. Let  $X$  be a Bernoulli variate with p.d.f.

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x}; x = 0, 1; 0 \leq \theta \leq 1.$$

Employ S.P.R.T. for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , and obtain its A.S.N. and O.C. functions.

[Delhi Univ. B.Sc. (Stat. Hons.), 1993, '85]

4. (a) Explain how the sequential test procedure differs from the Neyman-Pearson test procedure.

Develop the S.P.R.T. for testing  $H_0 : \pi = \pi_0$  against  $H_1 : \pi = \pi_1$ , based on a random sample from a binomial population with parameters  $(n, \pi)$ ,  $n$  being known. Obtain its O.C. and A.S.N. functions.

- (b) Obtain the sequential probability ratio test of the hypothesis  $H_0 : \theta = \frac{1}{3}$  against  $H_1 : \theta = \frac{2}{3}$  for the distribution :

$$f(x; \theta) = \begin{cases} \theta^x (1 - \theta)^{1-x}, & \text{for } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

[Madras Univ. B.Sc., 1988]

5. Develop S.P.R. test for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , ( $\theta_1 > \theta_0$ ), where  $\theta$  is the parameter of a Poisson distribution. Find approximate expressions for *OC* function and *ASN* function of the test.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

6. Describe S.P.R.T., its *OC* and *ASN* functions.

Construct S.P.R.T. for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , ( $0 < \theta_0 < \theta_1$ ), on the basis of a random sample drawn from the Pareto distribution with density function :

$$f(x, \theta) = \frac{\theta a^\theta}{x^{\theta+1}}, \quad x \geq a.$$

Also obtain its O.C. function and A.S.N. function.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

7. Explain how the sequential test procedure differs from the Neyman-Pearson test procedure.

Develop the S.P.R.T. for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1 (> \theta_0)$ , based on a random sample of size  $n$  from a population with p.d.f.

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}, x > 0, \theta > 0.$$

Also obtain its *A.S.N.* and *O.C.* functions.

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

8. Let  $X$  have the p.d.f.

$$f(x, \theta) = \begin{cases} \theta e^{-\theta x}; & x \geq 0, \theta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

For testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ , construct the *S.P.R.T.* and obtain its *ASN* and *OC* functions.

[Indian Civil Services (Main), 1989;  
Delhi Univ. B.Sc. (Stat. Hons.), 1982, 1986]

9. (a) What is a sequential test? How will you develop an optimum test of a specified strength for a simple null hypothesis versus a simple alternative?

(b) Find expressions for the sample size expected for termination of *SPRT* both under  $H_0$  and  $H_1$ . Clearly state all the assumptions made.

(c) A random variable follows the normal distribution  $N[\theta, \sigma^2]$ , where  $\sigma^2$  is known. Derive the *SPRT* for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ . Obtain the approximate expression for the *OC* function.

[Indian Civil Services (Main), 1990]

10. To test sequentially the hypothesis  $H_0$  that the distribution is given by  $P(X = -1) = P(X = 1) = P(X = 2) = \frac{1}{3}$  against the alternative  $H_1$  that it is given by  $P(X = -1) = P(X = 1) = \frac{1}{4}$ ;  $P(X = 2) = \frac{1}{2}$ , it is decided to continue sampling as long as  $-\frac{(n+1)}{2} < S_n < \frac{n+2}{2}$ , where  $S_n = X_1 + X_2 + \dots + X_n$ , the  $X_k$ 's being the successive observations. Compute the probability under  $H_0$  and under  $H_1$  that the procedure will terminate with the fourth observation or earlier.

11.  $X_1, X_2, \dots, X_n$  be a sequence of *i.i.d.* observations from  $N(\mu, \sigma^2)$ , where  $\mu$  is known,  $\sigma^2$  being unknown. Obtain the *SPRT* for testing  $H_0 : \sigma^2 = \sigma_0^2$ , against  $H_1 : \sigma^2 = \sigma_1^2 (> \sigma_0^2)$ . Also obtain its *OC* function and *A.S.N.* function.

### ADDITIONAL EXERCISE ON CHAPTER XVI

1. (a) "An examiner may pass a dull student or may fail a good student". Explain the above statement with reference to type-I and type-II errors.

2. A single value  $x$  is drawn from a normal population with mean  $m$  and variance 25. The null hypothesis  $H_0 : m = 50$  is accepted if  $x \leq 75$ , otherwise  $H_1 : m = 60$  is considered true. Evaluate the type I and type II errors.

3. Let  $p$  be the proportion of smokers in a certain city. You desire to test the hypothesis  $H_0 : p = \frac{1}{2}$  against  $H_1 : p = \frac{3}{4}$ . If you reject  $H_0$  when 60 persons or more are found smokers in a sample of 100 persons, compute the significance level and power of the test.

4. Let  $X_1, X_2, \dots, X_{20}$  be a random sample of size 20 from a Poisson distribution with mean  $\theta$ . Show that the critical region defined by  $\sum_{i=1}^{20} x_i \geq 5$ , is a uniformly most powerful critical region for testing  $H_0 : \theta = 1/10$  against  $H_1 : \theta > 1/10$ .

5. Let  $X_1, X_2, \dots, X_n$  denote a random sample from a normal distribution  $N(\theta, 16)$ . Find the sample size  $n$  and a uniformly most powerful test of  $H_0 : \theta = 25$  against  $H_1 : \theta < 25$ , with power function  $K(\theta)$  so that approximately  $K(25) = 0.10$  and  $K(23) = 0.90$ .

6. In testing  $H_0 : \sigma = \sigma_0$  against  $H_1 : \sigma = \sigma_1 (\neq \sigma_0)$ , for the distribution :

$$f(x) = \frac{1}{\sigma} \exp \left[ -\left( \frac{x - \theta}{\sigma} \right)^2 \right], (\theta \leq x < \infty, \sigma > 0)$$

Show that the UMP test is of the form

$$\sum x_i \geq \text{constant} \text{ and } \sum x_i \leq \text{constant.}$$

7.  $X_1, X_2$  is a random sample from a distribution with p.d.f.  $f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}$ ,  $x > 0, \theta > 0$ . The hypothesis  $H_0 : \theta = 2$  is tested against  $H_1 : \theta > 2$  and is rejected if and only if  $X_1 + X_2 \geq 9.5$ . Obtain the power function and the significance level of the test. Also find the probability of type II error when  $\theta = 4$ .

8. On the basis of a single observation  $x$  from the following p.d.f.

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} (x > 0; \theta > 0)$$

the null hypothesis,  $H_0 : \theta = 1$  against the alternative hypothesis  $H_1 : \theta = 4$ , is tested by using a set

$$C = \{x : x > 3\}$$

as the critical region. Prove that the critical region  $C$  provides a most powerful test of its size. What is the power of the test?

9. Let  $X$  be a single observation from the density  $f(x; \theta) = 2\theta x + 1 - \theta$ ,  $0 < x < 1, |\theta| \leq 1$ ; zero otherwise. Find the best critical region of size  $\alpha$ , for testing  $H_0 : \theta = 0$  against  $H_1 : \theta < 0$ . Express the power function of this test in terms of  $\alpha$ . Is the test uniformly most powerful? Explain.

10.  $X_1, X_2, \dots, X_n$  is a random sample of size  $n$  from  $N(\theta, 100)$ . For testing  $H_0 : \theta = 75$  against  $H_1 : \theta > 75$ , the following test procedure is proposed :

Reject  $H_0$  if  $\bar{x} \geq c$ ; Accept  $H_0$  if  $\bar{x} < c$ .

Determine  $n$  and  $c$  so that the power function  $P(\theta)$  of the test satisfies

$$P(75) = 0.159 \text{ and } P(77) = 0.841.$$

11. Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution having p.d.f.

$$f(x, \theta) = \frac{[x(1-x)]^{\theta-1}}{B(\theta, \theta)}, 0 < x < 1, \theta > 0 \\ = 0, \text{ elsewhere}$$

Show that the best critical region for testing  $H_0 : \theta = 1$  against  $H_1 : \theta = 2$  is

$$C = \left\{ (x_1, x_2, \dots, x_n) : c \leq \prod_{i=1}^n x_i (1 - x_i) \right\}.$$

12. Let  $X$  be a single observation from the distribution with p.d.f.

$$\begin{aligned} f(x; \theta) &= \theta e^{-\theta x}; \quad 0 < x < \infty, (\theta > 0) \\ &= 0, \quad \text{elsewhere.} \end{aligned}$$

Obtain the best critical region of size  $\alpha$  for testing  $H_0 : \theta = 1$  against  $H_1 : \theta = 2$ . Also obtain the power of this test.

[Delhi Univ. M.Sc. (Maths), 1990]

13. Let  $(x_1, x_2, \dots, x_9)$  be a random sample from  $N(\mu, 9)$ . To test the hypothesis  $H_0 : \mu = 40$  against  $H_1 : \mu \neq 40$ , consider the following two critical regions :

$$C_1 = \{\bar{x} : \bar{x} \geq a_1\}$$

$$C_2 = \{\bar{x} : |\bar{x} - 40| \geq a_2\}$$

(i) Obtain the values of  $a_1$  and  $a_2$  so that the size of each critical region is 0.05.

(ii) Calculate the power of the two critical regions when  $\mu = 39$  and  $\mu = 41$  and comment on the results.

[Delhi Univ. M.A. (Eco.), 1992]

14. (a) For a sample of size 25 from a normal population  $N(\mu, 25)$ ,  $\bar{X} = 11.5$ . Test the hypothesis  $H_0 : \mu = 10$  against the alternative  $H_1 : \mu > 10$ . Calculate the power of the test for  $\mu = 11$ .

[Delhi Univ. M.A. (Eco.), 1988]

(b) Let  $X \sim N(\mu, 25)$ . The null and alterantive hypotheses are :

$$H_0 : \mu = 10 \text{ and } H_1 : \mu < 10$$

(i) Give the best test of size  $\alpha = 0.05$  for a sample of size 25. (No derivation is expected).

(ii) Calculate the power of the test for  $\mu = 8$ .

[Delhi Univ. M.A. (Eco.), 1990]

15.  $X$  is normally distributed with  $\sigma = 5$  and it is desired to test  $H_0 : \mu = 105$  against  $H_1 : \mu = 110$ . How large a sample shoud be taken if the probability of accepting  $H_0$  when  $H_1$  is true is 0.02 and if a critical region of size 0.05 is used ?

(Agra Univ. B.Sc. 1989)

16. Let  $p$  be the probability that a given die shows an even number. To test  $H_0 : p = \frac{1}{2}$  against  $H_1 : p = \frac{1}{3}$ ; the following procedure is adopted. Toss the die twice and accept  $H_0$  if both times it shdws even number. Find the probabilities of type I and type II errors.

(Delhi Univ. M.C.A., 1990)

17. The p.d.f. of  $x$  is given by  $f(x) = \frac{1}{\theta}, 0 < x < \theta$ . Let the null hypothesis be  $H_0 : \theta = \frac{4}{3}$  against the alternative hypothesis  $H_A : \theta > \frac{4}{3}$ . We have a random sample of one observation. The critical region is defined by  $C = \{x : x > 1\}$ .

(i) Find the significance level of the test.

(ii) Find the power of the test for  $\theta = 7/3$  and  $\theta = 10/3$ .

[Delhi Univ. M.A. (Eco.), 1991]

**TABLE I**  
**LOGARITHMS**

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	-0000	0043	0086	0120	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	-0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	-1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	-1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	-1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	-2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	-2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	-2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	-2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	-3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	-3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	-3424	3444	3464	3483	3502	3522	3541	3562	3579	3598	2	4	6	8	10	12	14	15	17
23	-3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	-3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	-3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	-4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	-4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	-4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	-4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	-4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	-4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	-5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	-5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	-535	5315	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	-5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	-5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	-5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	-5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	-5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	-6021	6031	6042	6053	6065	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	-6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	-6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	-6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	-6435	6444	6454	6465	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	-6532	6542	6551	6561	6571	6380	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	-6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	-6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	6	7	8	
48	-6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	5	6	7	8	
49	-6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	
50	-6990	6993	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	-7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	-7160	7168	7177	7185	7193	7202	7210	7218	7225	7235	1	2	2	3	4	5	6	7	7
53	-7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	-7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7

**TABLE-I**  
**LOGARITHMS**

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
55	.7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	.742	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	.7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	.7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	.7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	.7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	.7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	.7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	.7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	.8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	.8129	8136	8142	8149	8156	8162	8169	8176	8183	8189	1	1	2	3	3	4	5	5	6
66	.8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	.8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	.8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	.8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	.8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	.8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	6
72	.8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	6
73	.8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	5	6	
74	.8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	6
75	.8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	6
76	.8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	6
77	.8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	.8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	.8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	.9331	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	.9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	.9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	.9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	.9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	.9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	.9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	.9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	.9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	.9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	.9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	.9590	9596	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	.9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	.9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	.9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	.9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	.9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	.9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	.9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	.9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4

**TABLE II**  
**ANTILOGARITHMS**

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
.02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
.03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	2	2	2
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	2	2	2	2	2
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	2	2	2	2	2
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	2	2	2	2	2
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
.09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
.21	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	1	1	1	2	2	2	3	3	3
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
.18	1514	1717	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
.20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
.22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	2	2	2	3	3	4
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	3	4
.28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	3	3	4	4	5
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	4	5
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	3	4	4	5
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	3	4	4	5
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	3	4	4	6
.43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	3	3	4	4	5	6
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
.45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
.48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	4	5	6	6	6
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1	1	2	3	4	4	5	6	6

**TABLE II**  
**ANTILOGARITHMS**

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
.52	3311	3319	3327	3334	3342	3550	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
.54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	3	3	4	5	6	6	7
.55	3548	3556	3563	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
.56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	3	4	5	6	7	8
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
.71	5129	5140	5152	5165	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
.72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
.75	5623	5636	5649	5662	5675	5889	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
.79	6166	6188	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
.82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
.85	7079	7096	7112	7129	7445	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7344	7261	7278	7295	7311	7328	7345	7362	7339	7396	2	3	5	7	8	10	12	13	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20

**TABLE II**  
**POWERS, ROOTS AND RECIPROCALS**

$n$	$n^2$	$n^3$	$\sqrt{n}$	$\frac{3}{\sqrt[n]{n}}$	$\sqrt[3]{10^n}$	$\frac{3}{\sqrt[3]{10^n}}$	$\frac{3}{\sqrt[3]{100^n}}$	$\frac{1}{\sqrt[n]{n}}$
1	1	1	1	1	3.162	2.154	4.642	1
2	4	8	1.414	1.260	4.772	2.714	5.848	.5000
3	9	27	1.732	1.442	5.477	3.107	6.694	.3333
4	16	64	2	1.587	6.325	3.420	7.368	.2500
5	25	125	2.236	1.710	7.671	3.684	7.937	.2000
6	36	216	2.449	1.817	7.745	3.915	8.434	.1667
7	49	343	2.646	1.913	8.361	4.121	8.879	.1429
8	64	512	2.828	2.000	8.944	4.309	9.283	.1250
9	81	729	3.000	2.080	9.487	4.481	9.655	.1111
10	100	1000	3.162	2.154	10.0	4.642	10.000	.1000
11	121	1331	3.317	2.224	10.488	4.791	10.323	.09091
12	144	1728	3.464	2.289	10.954	4.932	10.627	.08333
13	169	2197	3.606	2.351	11.402	5.066	10.914	.07692
14	196	2744	3.742	2.410	11.832	5.192	11.187	.07143
15	225	3375	3.873	2.466	12.247	5.313	11.447	.06667
16	250	4096	4.000	2.520	12.649	5.429	11.696	.06250
17	289	4913	4.123	2.571	13.038	5.540	11.935	.05882
18	324	5832	4.243	2.621	13.416	5.646	12.164	.05556
19	361	6859	4.359	2.568	13.784	5.749	12.386	.05263
20	400	8000	4.472	2.714	14.142	5.848	12.599	.05000
21	441	9261	4.583	2.759	14.491	5.944	12.806	.04762
22	484	10648	4.690	2.802	14.832	6.037	13.606	.04545
23	529	12167	4.796	2.844	15.166	6.127	13.200	.04167
24	576	13824	4.899	2.884	15.492	6.214	13.389	.04167
25	625	15625	5.000	2.924	15.811	6.300	13.572	.04000
26	676	17576	5.099	2.962	16.125	6.383	13.751	.03846
27	729	19683	5.196	3.000	16.432	6.463	13.925	.03704
28	784	21952	5.292	3.037	16.733	6.542	14.095	.03571
29	841	24389	5.385	3.072	17.029	6.619	14.260	.03448
30	900	27000	5.477	3.107	17.321	6.694	14.422	.03333
31	961	29791	5.568	3.141	17.607	6.768	14.581	.03226
32	1024	32768	5.657	3.175	17.889	6.9840	17.736	.03125
33	1089	35937	5.745	3.208	18.166	6.910	14.888	.03030
34	1156	39304	5.831	3.240	18.439	6.980	15.037	.02941
35	1225	42875	5.916	3.271	18.708	7.047	15.183	.02857
36	1296	46656	6.000	3.302	18.974	7.114	15.326	.02778
37	1369	50653	6.083	3.332	19.235	7.179	15.467	.02703
38	1444	54872	6.164	3.362	19.494	7.243	15.605	.02632
39	1521	59319	6.245	3.391	19.748	7.306	15.741	.02504
40	1600	64000	6.325	3.420	20.00	7.368	15.874	.0250
41	1681	68921	6.403	3.448	20.248	7.429	16.005	.02439
42	1764	74088	6.481	3.476	20.494	7.489	16.134	.02381
43	1849	79507	6.557	3.503	20.736	7.548	16.261	.02326
44	1936	85184	6.633	3.530	20.976	7.606	16.386	.02273
45	2025	91125	6.708	3.557	21.213	7.663	16.510	.02222
46	2116	97336	6.782	3.583	21.448	7.719	16.631	.02174
47	2209	103823	6.856	3.609	21.679	7.775	16.751	.02128
48	2304	110592	6.928	3.634	21.909	7.830	16.869	.02083
49	2401	117649	7.000	3.659	22.136	7.884	16.985	.02041
50	2500	125000	7.071	3.684	22.361	7.937	17.100	.020

**TABLE III**  
**POWERS, ROOTS AND RECIPROCALS**

$n$	$n^2$	$n^3$	$\sqrt{n}$	$\frac{3}{\sqrt{n}}$	$\sqrt{10n}$	$\frac{3}{\sqrt{10n}}$	$\frac{3}{\sqrt{100n}}$	$\frac{1}{n}$
51	2601	132651	7.141	3.708	22.583	7.990	17.213	.01961
52	2704	140608	7.211	3.733	22.804	8.041	17.325	.01923
53	2809	148877	7.280	3.756	23.022	8.093	17.435	.01887
54	2916	157464	7.348	3.780	23.238	8.143	17.544	.01852
55	3025	166375	7.416	3.803	23.452	8.193	17.652	.01818
56	3136	175616	7.483	3.832	23.664	8.243	17.758	.01786
57	3249	185193	7.550	3.849	23.875	8.291	17.863	.01754
58	3364	195112	7.616	3.871	24.083	8.340	17.967	.01724
59	3481	205379	7.681	3.893	24.290	8.387	18.070	.01695
60	3600	216000	7.746	3.915	24.495	8.334	18.171	.01667
61	3721	226981	7.810	3.936	24.698	8.481	18.272	.01639
62	3844	238328	7.874	3.958	24.900	8.527	18.371	.01613
63	3969	250047	7.937	3.979	25.100	8.573	18.469	.01587
64	4096	262144	8.000	4.000	25.298	8.618	18.566	.01562
65	4225	274625	8.062	4.021	25.495	8.662	18.663	.01538
66	4356	287496	8.124	4.041	25.690	8.707	18.758	.01515
67	4489	300763	8.185	4.062	25.884	8.750	18.852	.01493
68	4624	314432	8.246	4.082	26.077	8.794	18.945	.01471
69	4761	328509	8.307	4.102	26.268	8.837	19.038	.01449
70	4900	343000	8.367	4.121	26.458	8.879	19.129	.01429
71	5041	357911	8.426	4.141	26.646	8.921	19.220	.01408
72	5184	373248	8.485	4.160	26.833	8.963	19.310	.01389
73	5329	389017	8.544	4.179	27.019	9.004	19.399	.01370
74	5476	405224	8.602	4.198	27.203	9.045	19.487	.01351
75	5625	421875	8.660	4.217	27.38	9.086	19.574	.01333
76	5776	438976	8.718	4.236	27.568	9.126	19.661	.01316
77	5929	456533	8.775	4.254	27.740	9.166	19.747	.01299
78	6084	474552	8.832	4.273	27.928	9.205	19.832	.01282
79	6241	493039	8.888	4.291	28.107	9.244	19.916	.01266
80	6400	512000	8.944	4.309	28.284	9.283	20.000	.01250
81	6561	531441	9.000	4.327	28.460	9.322	20.083	.01235
82	6724	551368	9.055	4.344	28.636	9.360	20.165	.01220
83	6889	571787	9.110	4.362	28.810	9.398	20.247	.01205
84	7056	592704	9.165	4.380	28.983	9.435	20.328	.01190
85	7225	614125	9.220	4.397	29.155	9.473	20.408	.01176
86	7396	636056	9.274	4.414	29.326	9.510	20.488	.01163
87	7569	658503	9.327	4.431	29.496	9.546	20.507	.01149
88	7744	681472	9.381	4.448	29.665	9.583	20.646	.01136
89	7921	704969	9.434	4.465	29.833	9.619	20.224	.01124
90	8100	729000	9.487	4.487	30.000	9.655	20.801	.01111
91	8281	753571	9.539	4.498	30.166	9.691	20.878	.01099
92	8464	775688	9.592	4.514	30.332	9.726	20.954	.01087
93	8649	804357	9.644	4.531	30.496	9.761	21.029	.01075
94	8830	830584	9.695	4.547	30.659	9.796	21.105	.01064
95	9023	857375	9.747	4.563	30.822	9.830	21.179	.01053
96	9216	884736	9.798	4.579	30.984	9.865	21.253	.01042
97	9409	912673	9.849	4.595	31.145	9.899	21.327	.01031
98	9604	941192	9.899	4.610	31.305	9.933	21.400	.01020
99	9801	970299	9.900	4.626	31.464	9.967	21.472	.01010
100	10000	1000000	10.000	4.642	31.623	10.000	21.544	.01000

**TABLE IV**  
**AREAS UNDER NORMAL CURVE**

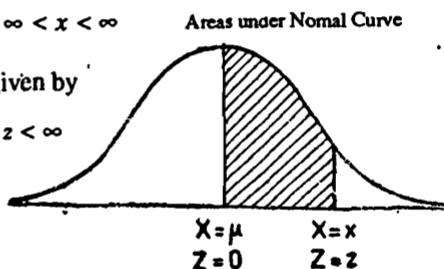
**Normal probability curve is given by**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right\} \quad -\infty < x < \infty$$

and standard normal probability curve is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right), -\infty < z < \infty$$

where  $Z = \frac{X - E(X)}{\sigma_X} \sim N(0, 1)$



The following table gives the shaded area in the diagram viz.,  $P(0 < Z < z)$  for different values of  $z$ .

## TABLE OF AREAS

**TABLE V**  
 ORDINATES OF THE NORMAL PROBABILITY CURVE

The following table gives the ordinates of the standard normal probability curve, i.e., it gives the value of

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2), -\infty < z < \infty$$

for different values of  $z$ , where

$$Z = \frac{X - E(X)}{\sigma_x} = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Obviously  $\phi(-z) = \phi(z)$ .

**TABLE VI**  
**SIGNIFICANT VALUES  $\chi^2(\alpha)$  OF CHI-SQUARE**  
**DISTRIBUTION. (RIGHT TAIL AREAS FOR GIVEN PROBABILITY  $\alpha$ ,**  
**where**

$$P = P_r (\chi^2 > \chi^2(\alpha)) = \alpha$$

AND  $v$  IS DEGREES OF FREEDOM (d.f.)

Degree of freedom (v)	Probability (Level of significance)						
	0 = .99	0.95	0.50	0.10	0.05	0.02	0.01
1	.000157	.00393	.455	2.706	3.841	5.214	6.635
2	.0201	.103	1.386	4.605	5.991	7.824	9.210
3	.115	.352	2.366	6.251	7.815	9.837	11.341
4	.297	.711	3.357	7.779	9.488	11.668	13.277
5	.554	1.145	4.351	9.236	11.070	13.388	15.086
6	.872	1.635	5.348	10.645	12.592	15.033	16.812
7	1.239	2.167	6.346	12.017	14.067	16.622	18.475
8	1.646	2.733	7.344	13.362	15.507	18.168	20.090
9	2.088	3.325	8.343	14.684	16.919	19.679	21.666
10	2.558	3.940	9.340	15.987	18.307	21.161	23.209
11	3.053	4.575	10.341	17.275	19.675	22.618	24.725
12	3.571	5.226	11.340	18.549	21.026	24.054	26.217
13	4.107	5.892	12.340	19.812	22.362	25.472	27.688
14	4.660	6.571	13.339	21.064	23.685	26.873	29.141
15	4.229	7.261	14.339	22.307	24.996	28.259	30.578
16	5.812	7.962	15.338	23.542	26.296	29.633	32.000
17	6.408	8.672	16.338	24.769	27.587	30.995	33.409
18	7.015	9.390	17.338	25.989	28.869	32.346	34.805
19	7.633	10.117	18.338	27.204	30.144	33.687	36.191
20	8.260	10.851	19.337	28.412	31.410	35.020	37.566
21	8.897	11.591	20.337	29.615	32.671	36.348	38.932
22	9.542	12.338	21.337	30.813	33.924	37.659	40.289
23	10.196	13.091	22.337	32.007	35.172	38.968	41.638
24	10.856	13.848	23.337	32.196	36.415	40.270	42.980
25	11.524	14.611	24.337	34.382	37.655	41.566	44.314
26	12.198	15.379	25.336	35.363	38.885	41.856	45.642
27	12.879	16.151	26.336	36.741	40.113	44.140	46.963
28	13.565	16.928	27.336	37.916	41.337	45.419	48.278
29	14.256	17.708	28.336	39.087	42.557	46.693	49.588
30	14.953	18.493	29.336	40.256	43.773	47.962	50.892

Note. For degrees of freedom (v) greater than 30, the quantity  $\sqrt{2\chi^2} - \sqrt{2v - 1}$  may be used as a normal variate with unit variance.

**TABLE VII**  
**SIGNIFICANT VALUES  $t_0(\alpha)$  OF  $t$ -DISTRIBUTION**  
**(TWO TAIL AREAS)**  
 $P[|t| > t_0(\alpha)] = \alpha$

$df,$ (v)	Probability (Level of Significance)					
	0.50	0.10	0.05	0.02	0.01	0.001
1	1.00	6.31	12.71	31.82	63.66	636.62
2	0.82	0.92	4.30	6.97	6.93	31.60
3	0.77	2.35	3.18	4.54	5.84	12.94
4	0.74	2.13	2.78	3.75	4.60	8.61
5	0.73	2.02	2.57	3.37	4.03	6.86
6	0.72	1.94	2.45	3.14	3.71	5.96
7	0.71	1.90	2.37	3.00	3.50	5.41
8	0.71	1.80	2.31	2.90	3.36	5.04
9	0.70	1.83	2.26	2.82	3.25	4.78
10	0.70	1.81	2.23	2.76	3.17	4.59
11	0.70	1.80	2.20	2.72	3.11	4.44
12	0.70	1.78	2.18	2.68	3.06	4.32
13	0.69	1.77	2.16	2.03	3.01	4.22
14	0.69	1.76	2.15	2.62	2.98	4.14
15	0.69	1.75	2.13	2.60	2.95	4.07
16	0.69	1.75	2.12	2.58	2.92	4.02
17	0.69	1.74	2.11	2.57	2.90	3.97
18	0.69	1.73	2.10	2.55	2.88	3.92
19	0.69	1.73	2.09	2.54	2.86	3.88
20	0.69	1.73	2.09	2.53	2.85	3.85
21	0.69	1.72	2.08	2.52	2.83	3.83
22	0.69	1.72	2.07	2.51	2.82	3.79
23	0.69	1.71	2.07	2.50	2.81	3.77
24	0.69	1.71	2.06	2.49	2.80	3.75
25	0.68	1.71	2.06	2.49	2.79	3.73
26	0.68	1.71	2.06	2.48	2.78	3.71
27	0.68	1.70	2.05	2.47	2.77	3.69
28	0.68	1.70	2.05	2.47	2.76	3.67
29	0.68	1.70	2.05	2.46	2.76	3.66
30	0.68	1.70	2.04	2.46	2.75	3.65
$\infty$	0.67	1.65	1.96	2.33	2.58	3.29

**TABLE VIII**  
**SIGNIFICANT VALUES OF THE VARIANCE-RATIO**  
**F-DISTRIBUTION (RIGHT-TAIL AREAS)**  
**5 PER CENT POINTS**

$v_1$	1	2	3	4	5	6	8	12	24	$\infty$
$v_2$										
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.35	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.65
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.46
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.78	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.865	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.365	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	4.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	5.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.54	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.4	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.96	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.76
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	4.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.60
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.30
120	3.92	3.87	2.68	2.45	2.29	2.17	2.02	1.83	1.62	1.25
240	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00



# Index

A		—Recurrence relation for moments	
Absolute moments	3·25	Bivariate frequency distribution	7·9
Additional theorem of		Bivariate normal distribution	10·32
—expectation	6·4	—Conditional distributions	10·84
—probability	4·30	—Marginal distributions	10·88
Additive property of		—M.g.f.	10·86
—Binomial variates	7·15	—moments	10·91
—Chi-square	13·7	Blackwellisation	15·34
—Gamma variates	8·70	Boole's inequality	4·33
—Normal variates	8·26	Borel Cantelli lemma	6·115
—Poisson variates	7·47	Buffon's needle problem	4·83
Algebra of events	4·21	C	
—sets	4·15	Cauchy distribution	8·98
Alternative hypothesis	12·6	Central limit theorem	8·105
Analysis of variance	14·67, 16·54, 16·49	—De Moivre's theorem	8·105
A posteriori probability	4·69	—Liaponouff's theorem	8·109
Approximate distributions .		—Lindberg-Levy theorem	8·107
—Binomial for hyper-geometric	7·91	Central tendency	2·6
—Normal for binomial	7·22	Characteristic function	6·77
—Normal for Chi-square	13·6	—multivariate	6·84
—Normal for gamma	8·69	—Properties	6·78
—Normal for Poisson	7·56	—Theorems	6·79
—Normal for t	14·14	—uniqueness theorem of	6·88
—Poisson for binomial.	7·40	Charlier's checks	3·24
—Poisson for negative binomial	7·75	Chebychev inequality	6·97
A priori probability	4·69	Chi-square distribution	13·1
Arithmetic mean	2·6	—non-central	13·69
—Demerits	2·11	Classical definition of probability	4·3
—Merits	2·10	Class frequency	11·1
—Properties	2·8	Class limits	2·3
—Weighted	2·11	Class sets	4·17
Array	2·1	—field	4·17
Association of attributes	11·15	—ring	4·17
Attributes	11·1	—σ-field	4·17
Average sample number	16·71	—σ-ring	4·17
A.S.N. function	16·71	Co-efficient of dispersion	3·12
Axiomatic definition of probability	4·17	—skewness	3·32
		—variation	3·12
B		Complete sufficient statistic	15·32
Bartlett's test	13·68	Composite family of distributions	15·31
Bayes' theorem	4·69	Composite hypothesis	16·2
Bernoulli distribution	7·1	Compound distribution	8·116
Bernoulli trials	7·1	—Binomial distribution	8·116
Bernoulli law of large numbers	6·103	—Poisson distribution	8·117
Best linear-unbiased estimator	15·14	Conditional distribution function	5·46
Beta distribution of		Conditional expectation	6·54
First kind	8·70	—probability	4·35
Second kind	8·72	Conditional variance	6·54
Binomial distribution	7·1	Confidence interval	15·82
—Additive property	7·15	—limits	15·82
—Characteristic function	7·16	—For difference of means	12·41, 16·47
—Cumulants	7·16	—For means	12·32, 16·42
—Factorial moments	7·11	—For proportions 'p'	12·13
—Mean deviation	7·11	—For variances	16·52, 16·55
—Mode	7·12		
—Moment Generating Function	7·14		

Consistent estimator	15·2	—Non-Central F	14·67
—Invariance property	15·3	—Non-Central t	14·43
—Sufficient conditions	15·3	—Normal	8·17
Consistent Data	11·8	—Order statistic	8·136
Contingency table	13	—Pareto	16·76
Continuous distribution function	5·32	—Pearsonian system	8·120
—frequency distribution	2·4	—Poisson	7·40
Continuous random variable	5·13	—Power series	7·101
Convergence in distribution (or law)	8·106	—Rectangular	8·1
Convergence in probability	6·100	—Student's t	14·1
Convolution	6·126	—t distribution	14·4
Correlation coefficient	10·1	—Triangular	8·10
—limits	10·2	—Uniform	8·1
—properties	10·3	—Weibull	8·90
Correlation of ranks	10·39		
Correlation ratio	10·76		E
Covariance	6·11		
Cramer-Rao's inequality	15·22	Efficient estimators,	15·7
Cramer's theorem	8·111	Elementary events	4·19
Critical region	16·3	Empirical probability	4·4
Cumulants	6·72	Equally likely events	4·3
Cumulative frequency	2·13	Error function (N.D.)	8·30
Curve fitting	9·1	Errors of first and second kind	16·4
Curvilinear regression	10·49	Estimation	15·1
		Estimator (unbiased)	15·2
		Event	4·19
		Exhaustive events	4·2
		Expectation	6·1
		—of continuous random variable	6·1
D		Exponential distribution	8·85
Deciles	2·26, 5·15		F
Degree of freedom	13·51	Factorial moments	3·24
Degenerate random variable	7·1	Factorization theorem (Neyman)	15·18
De-Moivre-Laplace theorem	8·105	Favourable events	4·2
Dichotomy	11·1	Fisher's Lemma	13·17
Discrete distribution function	6·7	—Z-distribution	14·69
Dispersion	3·1	—Z-transformation	14·71
Distribution discrete	7·1	Fourier inversion theorem	6·87
—Continuous	8·1	Frequency	2·2
Distribution function	5·7	Frequency distribution	2·1
Distribution function of joint distribution	5·42	—polygon	2·5
Distributions :		—table	2·2
—Bernoulli	7·1	F-distribution (Snedecor's)	14·45
—Beta	8·70	—non-central	14·67
—Binomial	7·1		G
—Bivariate normal	10·84	Gama Distribution	8·68
—Cauchy	8·98	—Additive property	8·70
—Chi-square	13·1	—Cumulant Gen. Function	8·68
—Discrete uniform	8·1	—Moment Gen. Function	8·68
—Double exponential	8·89	Geometric distribution	7·83
—Exponential	8·85	Geometric mean	2·22
—F-distribution	14·45	Geometrical Probability	4·80
—Gamma	8·68	Goodness of fit	13·39
—Generalised power series	7·103	Grouped frequency distribution	2·2
—Geometric	7·83		
—Hypergeometric	7·88		
—Laplace	8·89		
—Logistic	8·92		
—Log normal	8·65		
—Multinomial	7·95		
—Negative binomial	7·72		
—Non-central Chi-square	13·69		

	H		—Dispersion	3·1
Harmonic mean	2·25		—Kurtosis	3·35
Hazoor Bazar's Theorem	15·54		—Skewness	3·32
Helly Bray Theorem	6·90	Median	2·13	
Histogram	2·4	—Derivation	2·19	
Hypergeometric distribution	7·88	—Demerits	2·16	
		—Merits	2·16	
	I	Median test	16·64	
Impossible event	4·19	Mesokurtic curve	3·35	
Independence of attributes	11·12	Methods of estimation	15·52	
Independent events	4·37	—Least square	15·73	
Intraclass correlation	10·81	—Maximum Likelihood	15·52	
Interval Estimation	15·82	—Minimum variance	15·69	
Inversion Theorem	6·87	—Moments	15·69	
	J	Mode	2·17	
Joint density function	5·44	—Derivation	2·19	
—Probability law	5·41	—Demerits	2·22	
—Probability distribution		—Merits	2·22	
function	5·42	Moments	3·21	
—Probability mass function	5·41	—Absolute	3·25	
	K	—Factorial	3·24	
Khintchin's Theorem	6·104	—Sheppard correction	3·23	
Koopman's Form	15·19	Moment generating function	6·67	
Kurtosis	3·35	—of binomial distribution	7·14	
	L	—of bivariate normal		
Laplace double exponential distribution	8·89	distribution	10·86	
Large sample tests	12·10	—of negative binomial		
Least square principle	9·1	distribution	7·74	
Leptokurtic	3·35	—of chi-square distribution	13·5	
Level of significance	12·7	—of exponential distribution	8·86	
Liapounoff's theorem	8·109	—of gamma distribution	8·68	
Likelihood ratio test	16·34	—of geometric distribution	7·85	
Linear transformation	13·16	—of multinomial distribution	7·96	
Lindeberg-Levy theorem	8·107	—of non-central chi-square		
Lines of regression	10·49	distribution	13·70	
Log-normal distribution	8·65	—of normal distribution	8·23	
	M	—of Poisson distribution	7·47	
Markoff's theorem	6·104	Moments of bivariate probability		
Mann Whitney U-test	14·62	distribution	6·54	
Marginal distribution function	5·43	Most efficient estimator	15·8	
—probability function		Multinomial distribution	7·95	
Mathematical expectation	6·1	Multiple correlation		
—of continuous r.v.	6·2	—coefficient	10·111	
Mean deviation	3·2	Multiple regression	10·105	
Mean square deviation	3·2	Multiplication law of probability	4·35	
Measures of		—expectation	6·6	
—Central tendency	2·6	Multivariate Characteristic		
—Correlation	10·1	Function	6·84	
—Correlation ratio	10·76	Mutually exclusive events	4·2	
	N	MVB estimator	15·24, 15·26	
		Negative binomial distribution	7·72	
		—Cumulants	7·74	
		—Moment Generating Function	7·74	
		—Probability Generating		
		Function	7·76	
		Neyman and Pearson Lemma	16·7	
		Non-parametric methods	16·59	

Normal Distribution	8 · 17	—Geometric	4 · 8
—Characteristics	8 · 20	—History	4 · 1
—Cumulant Generating Function	3 · 24	—Multiplicative Law	4 · 3
—Importance	8 · 31	Probability density function	5 · 1
—Mean Deviation	8 · 28	Probability distribution	5 · 1
—Median	8 · 23	—conditional	4 · 3
—Mode	8 · 22	Probability generating function	6 · 12
—Moments	8 · 24	—mass function	5 · 1
—Moment Generating Function	8 · 23	Probable error	10 · 3
Normal Curve	8 · 29	Purposive sampling	12 · 1
—equations	9 · 2		
Null hypothesis	12 · 6		
<b>Q</b>			
		Quartile	2 · 26, 5 · 1
		—deviation	3 · 1
<b>O</b>			
Ogive	2 · 27		
Operating characteristic(O.C.) curve	16 · 21		
O · C. function	16 · 71		
Order statistics	8 · 136	Random experiments	4 · 1
—Distribution function of $X(r)$	8 · 136	—sampling	12 · 1
—Joint p.d.f. of $X(r), X(s)$	8 · 138	Random variables	5 · 1
—p.d.f. of $X(1)$	8 · 137	—discrete	5 · 1
—p.d.f. of $X(n)$	8 · 137	Range	3 · 1
—p.d.f. of Range	8 · 140	Rank Correlation	10 · 3
<b>P</b>			
Paired t-test		Rao-Cramer inequality	15 · 2
Pairwise independent events	4 · 39	Rao-Blackwell theorem	15 · 3
Parameter	12 · 3	Rectangular distribution	8 · 1
Parameter space	15 · 1	Regression coefficients	10 · 5
Pareto distribution	16 · 76	Regression (linear and nonlinear)	10 · 5
Partial correlation	10 · 103	—curve	10 · 6
—Coefficient	10 · 114	—plane	10 · 10
—regression coefficient	10 · 105	Relation between t and F	14 · 6
Partition values	2 · 26	F and $\chi^2$	14 · 6
—Graphical location	2 · 27	Reproductive property (see additive property)	
Pearson $\beta$ and $\gamma$ coefficients	3 · 24	Residual variance (regression)	10 · 10
Pearson's distribution	8 · 120	Root mean square deviation	3 · 1
Percentiles	2 · 26	Run	16 · 6
Platykurtic	3 · 35	Run Test (Wald-Wolfowitz)	16 · 6
Poisson distribution	7 · 40		
—Additive Property	7 · 47	<b>S</b>	
—Characteristic function	7 · 47	Sample correlation coefficient	14 · 3
—Cumulants	7 · 47	Sample partial correlation coefficient	14 · 3
—Mode	7 · 44	Sample multiple correlation	
—Moments	7 · 47	coefficient	14 · 3
—Moment Generating Function	7 · 47	Sample space	4 · 18
—Recurrence relation for Moments	7 · 46	Sample standard deviation	
Poisson Process	7 · 42	—variance and covariance	12 · 29
Power of test	16 · 5	Sampling	12 · 1
Power series distribution	7 · 101	—attributes	12 · 11
Probability	4 · 1	—variables	12 · 20
—Addition law	4 · 30	Sampling distribution	12 · 3
—Axiomatic	4 · 17	Scatter diagram	10 · 1
—Bayes' Theorem	4 · 69	Semi-interquartile range	3 · 1
—Definitions of Terms	4 · 2	Sequential analysis	16 · 69
—Empirical	4 · 4	—probability ratio test	16 · 69
—Function	5 · 6	—A.S.N. function	16 · 71
		—O.C. Function	16 · 71

<b>Set</b>			
—complement	4·14	non-central	14·43
—disjoint	4·15	Transformation of	
—empty	4·15	—one-dimensional r.v.	5·70
—equal	4·14	—two-dimensional r.v.	5·73
<b>Sign test</b>	4·14	Triangular distribution	8·10
<b>Simple sampling</b>	16·65	Truncated distributions	8·151
<b>Simple hypothesis</b>	12·2	—Binomial Examples	8·54
<b>Skewness</b>	3·32	—Cauchy	8·55
<b>Spearman's Rank Correlation</b>	10·39	—Gamma	8·156
Coefficient	10·40	—Normal Examples	8·153
—Tied ranks		—Poisson	8·154, 8·155
<b>Standard deviation</b>	3·2	Type of Sampling	
—error	12·4		<b>U</b>
<b>Statistic</b>	12·3		
Statistics, meaning of	1·1	U-Test	16·66
Statistical probability	4·4	UMPT	16·7
<b>Stochastic Independence</b>	4·39	Unbiasedness of estimator	15·2
<b>Stratified sampling</b>	12·3	Uniform distribution (Discrete)	8·1
<b>Student's t-distribution</b>	14·2	—continuous	
<b>Sufficient statistics</b>	15·18	Uniqueness theorem of moment	
—Complete	15·31	generating functions	6·72
—Factorisation theorem	15·18		<b>V</b>
<b>T</b>			
<b>Tchebycheff inequality</b>		Variance	8·3
(see Chebychev's inequality)		Variance of residual	10·110
<b>Testing of hypothesis</b>	16·2		<b>W</b>
—composite hypothesis	16·1		
—critical region	16·3	Wald-Wolfowitz Test	16·61
—distribution free methods	16·59	Wald's SPRT	16·69
—level of significance	16·5	Weak law of large numbers	6·101
—most powerful test (MPT)	16·6	Weighted mean	2·11
—null hypothesis	12·6		<b>Y</b>
—Neyman-Pearson Lemma	16·7		
—non-parametric tests	16·59	Yates' continuity correction	13·57
—sequential test	16·69	Yule's coefficient of association	11·17
(see sequential analysis)		—colligation	11·17
—two kinds of errors	16·4	Yule's notation (multiple and	
—uniformly most powerful test	16·7	partial correlation)	10·104
(UMPT)			<b>Z</b>
<b>Test for randomness</b>			
—of significance	16·6		
<b>Tippet random numbers</b>	12·2		
<b>t-distribution (Student's)</b>	14·1	Zero-one law	6·117
Fisher	14·3		