26. A speak truth in 75% cases and B in 80% cases. In what percentage of cases are they likely to contradict each other in stating the same fact? **[Ans.** 35%]

27. A purse contains 2 silver and 4 copper coins. A second purse contains 4 silver and 3 copper. If a coin is pulled out a random from one of the two purses, what is the probability that is it a silver coin?

28. A student takes his examination in four subjects P, Q, R, S. He estimates his chances of passing in P as $\frac{4}{5}$, in Q as $\frac{3}{4}$, in R as $\frac{5}{6}$ and in S as $\frac{2}{3}$. To qualify, he must pass in P atleast two other subjects. What is the probability that the qualifies? $\left[\textbf{Ans.}\frac{61}{90}\right]$

29. Define a random experiment, sample space, event and mutually exclusive events. Give example of each.

1.4. CONDITIONAL PROBABILITY

Let A and B be two events associated with the same sample space of a random experiment. Then the probability of occurrence of A under the condition that B has already occurred, at $P(B) \neq 0$, is called conditional probability, denoted by $P(A/B)$.

We define, $$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{n(A \cap B)}{n(B)}, \quad \text{where } P(B) \neq 0 \text{ and}$$

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

$$= \frac{P(A \cap B)}{P(A)} = \frac{n(A \cap B)}{n(A)}, \text{ where, } P(A) \neq 0$$

1.4.1. Properties of Conditional Probability

Let A and B be events of a sample space S of an experiment, then we have

Property 1. $P(S/B) = P(B/B) = 1$

We know that $$P(S/B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

Also $$P(B/B) = \frac{P(B \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

Thus, $$P(S/B) = P(B/B) = 1$$

Property 2. If A and B are any two events of a sample space S and F is any event of S such that $P(F) \neq 0$, then

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

In particular, if A and B are disjoint events, then

$$P(A \cup B)/F) = \frac{P[(A \cup B) \cap F]}{P(F)}$$

$$= \frac{P[(A \cap F) \cup (B \cap F)]}{P(F)}$$

(by distributive law of union of set over intersection)

$$= \frac{P(A \cap F) + P(B \cap F) - P[(A \cap B) \cap F]}{P(F)}$$

$$= P(A/F) + P(B/F) - P((A \cap B)/F)$$

When A and B are disjoint events, then

$$P((A \cap B)|F) = 0$$

$$\Rightarrow \quad P((A \cup B) \backslash F) = P(A/F) + P(B/F)$$

Property 3. $P(E'/F) = 1 - P(E/F)$

From property 1, we know that

$$P(S/F) = 1$$

$$\Rightarrow \quad P(E \cup E'/F) = 1$$

$$\Rightarrow \quad P(E/F) + P(E'/F) = 1$$

Since E and E' are disjoint events

Thus $P(E'/F) = 1 - P(E/F)$

SOLVED EXAMPLES

Example 1.26. A pair of dice is rolled, find $P(A/B)$ if

A: 2 appears on atleast one dice.

B: sum of numbers appearing on dice is 6.

Solution. We have

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

$$B = \{1, 5)\ (2, 4)\ (3, 3)\ (4, 2)\ (5, 1)$$

$$A \cap B = \{(2, 4)\ (4, 2)\}$$

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

$$P(B) = \frac{5}{36}$$

Therefore,

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

Example 1.27. Two marbles are drawn successively from a box containing 3 black and 4 white marbles. Find the probability that both marbles are black if the first marble is not required before the second drawing.

Solution. Let B_1 is the event of drawing the first black marble.

Then, $P(B_1) = \frac{3}{7}$

Let B_2 be the event that the second marble drawn in black. Then

$P(B_2|B_1)$ = Conditional probability of the event B_2 given that B_1 has occurred

$$= \frac{2}{6}$$

Hence by multiplication rule, we get

$$P(B_1 \text{ and } B_2) = P(B_1 \cap B_2) = P(B_1)\, P(B_2|B_1)$$
$$= \frac{3}{7} \times \frac{2}{6} = \frac{1}{7}$$

Example 1.28. A card is drawn from a well shuffled deck of 52 cards and then second card is drawn, find the probability that the first card is a spade and then second card is a club if the first card is not replaced.

Solution. We have

$$P(\text{first card spade}) = P(S) = \frac{13}{52} = \frac{1}{4}$$

After the event of drawing a spade the deck has 51 cards 13 of which are clubs (C)

Therefore, $P(C|S) = \frac{13}{51}$

Hence, $P(S \text{ and } C) = P(S)\, P(C|S)$

$$= \frac{1}{4} \cdot \frac{13}{51} = \frac{13}{204}$$

Example 1.29. If $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{4}$, $P(A \cup B) = \frac{1}{2}$ determine : (i) $P(B|A)$ (ii) $P(A|B')$.

Solution. Given that: $P(A) = \frac{1}{3}, P(B) = \frac{1}{4}, P(A \cup B) = \frac{1}{2}$ gives $P(B') = \frac{3}{4}$,

From the addition theorem on the probability,

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

$\therefore$ $$P(A \cap B) = \frac{1}{3} + \frac{1}{4} - \frac{1}{2} = \frac{1}{12}$$

(i) $$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{12}}{\frac{1}{3}} = \frac{1}{4}$$

(ii) $$P(A \cap B') = P(A) - P(A \cap B)$$

(*i*) $$\frac{P(A \cap B')}{P(B')} = \frac{P(A)}{P(B')} - \frac{P(A \cap B')}{P(B')}$$

$$\Rightarrow \quad P(A/B') = \frac{\frac{1}{3}}{1-\frac{1}{4}} - \frac{\frac{1}{12}}{1-\frac{1}{4}} = \frac{\frac{1}{3}}{\frac{3}{4}} - \frac{\frac{1}{12}}{\frac{3}{4}}$$

$$= \frac{4}{3}\left[\frac{1}{3} - \frac{1}{12}\right] = \frac{4}{3}\left(\frac{4-3}{12}\right) = \frac{1}{3}$$

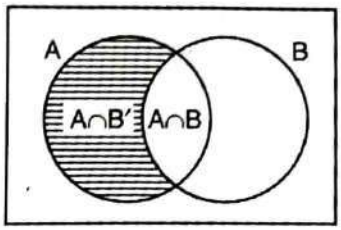

Fig. 1.1.

$$\therefore \quad P(A/B') = \frac{1}{3}$$

Example 1.30. A dice is thrown twice and the sum of the numbers appearing is observed to be 6. What is the conditional probability that the number 4 has appeared at least once ?

Solution. Consider the events.

A = number 4 appears atleast once

B = the sum of the number appearing is 6

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

and $$B = \{(1, 5) (2, 4) (3, 3) (4, 2) (5, 1)\}$$

$$P(A \cap B) = \{(2, 4) \ (4, 2)\}$$

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

Example 1.31. A market survey was conducted in four cities to find out the preference for brand *A* soap. The responses are shown below:

	Delhi	*Kolkata*	*Chennai*	*Mumbai*
Yes	**45**	**55**	**60**	**50**
No	**35**	**45**	**35**	**45**
No opinion	**5**	**5**	**5**	**5**

(*a*) What is the probability that a consumer preferred brand *A*, given that he was from Chennai?

(*b*) Given that a consumer preferred brand *A*, what is the probability that he was from Mumbai?

Solution. The information from responses during market survey is as follows:

	Delhi	*Kolkata*	*Chennai*	*Mumbai*	*Total*
Yes	45	55	60	50	210
No	35	45	35	45	160
No opinion	5	5	5	5	20
Total	85	105	100	100	390

Let X denote the event that a consumer selected at random preferred brand A. Then :

(*a*) The probability of a consumer preferred brand A, given that he was from Chennai :

$$P(X|C) = \frac{P(X \cap C)}{P(C)} = \frac{\frac{60}{390}}{\frac{100}{390}} = \frac{3}{5}$$

(*b*)The probability that the consumer belongs to Mumbai, given that he preferred brand A:

$$P(M/X) = \frac{P(M \cap X)}{P(X)} = \frac{\frac{50}{390}}{\frac{210}{390}} = \frac{5}{21}$$

Example 1.32. Data on the readership of a certain magazine show that the proportion of male readers under 35 is 0.40 and over 35 is 0.20. If the proportion of readers under is 0.70, find the proportion of subscribers that are 'females over 35 years'. Also calculate the probability that a randomly selected male subscriber is under 35 years of age.

Solution. Let us define the following events:

A : Reader of the magazine is a male.

B : Reader of the magazine is over 35 years of age.

Then in usual notations, we are given :

(*i*) The proporation of subscribers that are females over 35 years is :

$$P(A \cap B) = 0.20, P(A \cap \bar{B}) = 0.40$$

and $$P(\bar{B}) = 0.70 \Rightarrow P(B) = 0.30$$

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

$$= 0.30 - 0.20 = 0.10$$

(*ii*) The probability that a randomly selected male subscriber is under 35 years is :

$$P(\bar{B} \mid A) = \frac{P(A \cap \bar{B})}{P(A)} = \frac{0.40}{0.60} = \frac{2}{3}$$

$$[\because P(A) = P(A \cap B) + (A \cap \bar{B}) = 0.20 + 0.40 = 0.60]$$

Example 1.33. If the probability that a communication system will have high fidelity is 0.81 and the probability that it will have high fidelity and selectivity is 0.18, what is the probability that a system with high fidelity will also have selectivity?

Solution. Let A be the event that represent a communication system will have high fidelity

$\therefore \quad P(A) = 0.81$

Let $(A \cap B)$ be the event that represents high fidelity and selectivity.

$\therefore \quad P(A \cap B) = 0.18$

$\therefore$ The probability that a system will have high fidelity will also high selectivity (by using conditional probability) is

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.18}{0.81} = \frac{2}{9}$$

Example 1.34. A couple has two children. Find the probability that both children are boys, if it is known that at least one of the children is a boy.

Solution. Let B_i and G_i stands for i^{th} child be a boy and girl respectively. Then sample space can be expressed as

$$S = \{B_1 B_2, B_1 G_2, G_1 B_2, G_1 G_2\}$$

Consider the following events

$A =$ both the children are boys

$B =$ at least one of the children is a boy

Then $A = \{B_1 B_2\}$

$B = \{B_1 G_2, G_1 B_2, B_1 B_2\}$

So $A \cap B = \{B_1 B_2\}$

Required $$P(B|A) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

Example 1.35. The probability that a student selected at random from a class will pass in a Mathematics is $\frac{4}{5}$ and the probability that he/she passes in Mathematics and Computer Science is $\frac{1}{2}$. What is the probability that he/she will pass in computer science, if it is known that he has passed in mathematics?

Solution. Probability (Pass in Mathematics)

$$= \frac{4}{5} = P(M)$$

Probability (Passes in Mathematics and Computer Science)

$$= \frac{1}{2} = P(M \cap C)$$

$$P(C) = ?$$

(c) Required probability is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$= \frac{25}{100} + \frac{15}{100} - \frac{10}{100} = \frac{30}{100} = \frac{3}{10}$$

Example 1.40. A bag contains 6 white and 9 black balls. Four balls are drawn at a time. Find the probability for the first draw to give four white and second draw to give four black balls in each of the following cases:

(a) The balls are replaced before the second draw.

(b) The balls are not replaced before the second draw.

Solution. A : Getting 4 white ball in first draw

B : Getting 4 black ball in second draw

(a)
$$P(A) = \frac{{}^6C_4}{{}^{15}C_4} = \frac{6\times5\times4\times3}{15\times14\times13\times12} = \frac{360}{32760}$$

$$P(B) = \frac{{}^9C_4}{{}^{15}C_4} = \frac{9\times8\times7\times6}{15\times14\times13\times12} = \frac{3024}{32760}$$

Required probability $= P(A)\,P(B)$

$$= \frac{360}{32760}\times\frac{3024}{32760} = \frac{6}{5915}$$

(b)
$$P(A) = \frac{{}^6C_4}{{}^{15}C_4} = \frac{6\times5\times4\times3}{15\times14\times13\times12} = \frac{360}{32760}$$

$$P(B \mid A) = \frac{{}^9C_4}{{}^{11}C_4} = \frac{9\times8\times7\times6}{11\times10\times9\times8} = \frac{3024}{7920}$$

Required probability $= P(A)\,P(B|A)$

$$= \frac{360}{3270}\times\frac{3024}{7920} = \frac{3}{715}$$

EXERCISE 1.2

1. If $P(A) = \frac{6}{11}, P(B) = \frac{5}{11}, P(A \cup B) = \frac{7}{11}, P(A \cup B) = \frac{7}{11}$

(i) $P(A \cup B)$ (ii) $P\left(\frac{A}{B}\right)$ (iii) $P\left(\frac{B}{A}\right)$ [**Ans.** (i) $\frac{4}{11}$, (ii) $\frac{4}{5}$, (iii) $\frac{2}{3}$]

2. If $P(A) = \frac{3}{8}$, $P(B) = \frac{1}{2}$ and $P(A \cap B) = \frac{1}{3}$, find $P\left(\frac{\overline{A}}{B}\right)$ and $P\left(\frac{\overline{B}}{A}\right)$ [**Ans.** $\frac{3}{4}, \frac{3}{5}$]

3. A pair of dice is thrown. Let E be the event that sum is greater than or equal to 10 and F be the event the 5 appears on the first dice. Find $P(E/F)$. $\left[\text{Ans.}\frac{1}{3}\right]$

4. A pair of dice is thrown. If the two numbers appearing on them are different find the probability that
   (*i*) the sum of numbers is 6.
   (*ii*) the sum of number of 4 or less. $\left[\text{Ans. } (i)\frac{30}{36},\ (ii)\frac{2}{15}\right]$

5. A bag contains 10 white and 15 black balls. Two balls are drawn is succession without replacement. What is the probability that first is white and second is black. $\left[\text{Ans.}\frac{1}{4}\right]$

6. Find the probability of drawing a diamond card in each of the two consecutive draws from a well shuffled pack of cards. If the card drawn in not replaced after the first draw.
$\left[\text{Ans.}\frac{1}{17}\right]$

7. Two dice are thrown that it is known that first dice shows a six. Find the probability that the sum of numbers showing on the dice is 7.

8. A coin in tossed then a dice is thrown. Find the probability of a 6 given that heads came up.
$\left[\text{Ans.}\frac{1}{6}\right]$

9. The probability that a certain person will buy a shirt is 0.2 the probability that he will buy a trouser is 0.3 and the probability that he will buy a shirt given that he buys a trouser is 0.4. Find the probability that he will buy both shirt and trouser. Find also the probability that he will buy a trouser given that the buys a shirt. [**Ans.** 0.06]

10. A bag contain 10 white and 15 black balls. Two balls are drawn in succession without replacement. What is the probability that first is white and second is black. $\left[\text{Ans.}\frac{1}{4}\right]$

11. A dice is rolled twice and the sum of the numbers appearing on them is observed to be 6. What is the conditional probability that the number 4 has appeared at least once. $\left[\text{Ans.}\frac{2}{5}\right]$

12. A dice is rolled twice and the sum of the numbers appearing on them is observed to be 7. What is the conditional probability that the number 2 has appeared at least once. $\left[\text{Ans.}\frac{1}{3}\right]$

13. Find the probability of drawing a diamond card in each of the two consecutive draws from a well shuffled pack of cards. If the card drawn in not replaced after the first draw. $\left[\text{Ans.}\frac{1}{17}\right]$

14. A bag contains 5 white, 7 red and 8 black balls. If four balls are drawn one by one without replacement find the probability of getting all white balls. $\left[\text{Ans.}\frac{1}{969}\right]$

15. An urn contains 5 white and 8 black balls two successive drawing of three balls at a time are made such that the balls are not replaced before the second draw. Find the probability that the first draw gives 3 white balls and second draw gives 3 black balls.

16. A coin is tossed twice and the four possible outcome are assumed to be equally likely. If E is the event "both head and tail have occurred" and F the event "at most one tail has occurred". Find $P(E)$ $P(F)$ $P(E/F)$ and $P(F/E)$.

17. An urn contains 10 white and 3 back balls while another urn contains 3 white and 5 black balls. Two are drawn from first urn and put into the second urn and then a ball is drawn from the latter. Fid the probability that it is white ball. $\left[\text{Ans.}\frac{59}{130}\right]$

18. There are two bags containing 5 red, 7 white and 3 red, 12 white balls respectively. A ball is drawn from one of the two bags, find the probability of drawing a red ball. $\left[\text{Ans.}\frac{37}{120}\right]$

19. A purse contains 2 silver and 4 copper coins. A second purse contains 4 silver and 3 copper coins if a coin is pulled out at random from one of the two purses. What is the probability that a silver coin. $\left[\text{Ans.}\frac{19}{42}\right]$

20. A bag contains 6 red and 8 black balls and another bag contains 8 red and 6 black balls, A ball is drawn from first bag and without noticing it colour is put in the second bag. A ball is then drawn from the second bag. Find the probability that the ball drawn is red in colour. $\left[\text{Ans.}\frac{35}{108}\right]$

1.5. RANDOM VARIABLE AND PROBABILITY DISTRIBUTION

If an experiment is conducted under identical conditions, values so obtained may not be similar. Observations are always taken about a **factor** or **character** under study, which can take different values. This **factor** or **character** is termed as **variable.** The observations may be the number of certain objects or items or their measurements. These observations very even though the experiement is conducted under identical conditions. Hence, we have a set of outcomes of a random experiment. A **rule** that assigns a real number to each outcome is called **random variable.** The rule is nothing but a function of the variable, say, X that assigns a unique value to each outcome of the random experiment. It is clear that there is a value for each outcome, which it takes with certain probability. Thus when a variable X takes the value x_i with probability p_i, $(i = 1, 2, 3, \ldots n)$, then X is called **random variable** or **stochastic variable** or a **variate.**

1.6. DISCRETE RANDAM VARIABLE

A random variable X, which can take only a finite number of values in an interval of the domain is called discrete random variable.

Example:

1. Number appearing on top of a die when it is thrown.
2. The number of telephone calls received per day.
3. Number of mistakes in a page.
4. Number of defective items in a lot.

1.7. DISCRETE PROBABILITY DISTRIBUTION

If a random variable x can assume a discrete set of values say $x_1, x_2, \ldots x_n$ with respect to probabilities $p_1, p_2, \ldots p_n$ such that $p_1 + p_2 + \ldots p_n = 1$ *i.e.*, $\sum_{i=1}^{n} pi = 1$ then occurrences of values x_i with respective probabilities p_i is called the discrete probability distribution of X.

For example, In a throw of a pair of dice the sum (X) is discrete random variable which is an integer between 2 and 12 with probabilities $P(X)$ given as

X	2	3	4	5	6	7	8	9	10	11	12
$P(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

This constitute a discrete probability distribution.

1.8. PROBABILITY FUNCTION OR PROBABILITY MASS FUNCTION (pmf)

Probability function of probability mass function (p.m.f.) of a random variable X is mathematical function $p(x)$ which gives the probabilities corresponding to different possible discrete set of values say $x_1, x_2, x_3 \ldots\ldots x_n$ of variable x.

i.e., $$p(x_i) = p(x = x_i)$$

= probability that and on variable x assumes value x_i

The function $p(x)$ satisfies the condition.

(*i*) $p(x_i) \geq 0$

(*ii*) $\Sigma p(x_i) = 1$

1.9. CUMULATIVE DISTRIBUTION FUNCTION (DISTRIBUTION FUNCTION)

If X is a random variable then $P(X \leq x)$ is called the cumulative distribution function (cdf) or distribution function and is denoted by $F(x)$

$\therefore$ $$F(x) = p(X < x)$$

1.10. EXPECTATION OF A DISCRETE RANDOM VARIABLE

If x is discrete random variable which assumes the discrete set of values $x_1, x_2, \ldots x_n$ with respective probabilities $p_1, p_2 \ldots p_n$ then the expectation or expected value of x is denoted by $E(X)$ and defined as

$$E(X) = x_1 p_1 + x_2 p_2 + x_3 p_3 \ldots + x_n p_n$$

$$= \sum_{i=1}^{n} x_i p_i$$

Similarly the expected value of X^2 is defined as $E(X^2) = \sum_{i=1}^{n} x_i^2 p_i$

Properties:

1. If X is a random variable and 'a' is constant then
   (*i*) $E(a) = a$
   (*ii*) $E(aX) = a\,E(X)$
   (*iii*) $E(X - \mu) = 0$
2. If x and y are two random variables then $E(X \pm Y) = E(X) \pm E(Y)$.
3. $E(XY) = E(X)\,E(Y)$ if X and Y are two independent random variable.
4. if $y = ax + b$ where a and b are constant then $E(Y) = aE(X) + b$.

1.11. VARIANCE AND STANDARD DEVIATION OF DISCRETE RANDOM VARIABLE

The variable of discrete random variable X is expected value of $(X-\mu)^2$ where μ is mean of varaible X

$$Var\ X = V(x) = E(X-\mu)^2$$

$$= \sum_{i=1}^{n} (x_i - \mu)^2$$

$$= \Sigma p\,(x-\mu)^2$$

$$= \Sigma px^2 + \Sigma p\mu^2 - 2\Sigma p\mu x$$

$$= \Sigma px^2 + \mu^2\,\Sigma p - 2\mu\Sigma px$$

$$= \Sigma px^2 + \mu^2 - 2\mu^2$$

$$= \Sigma px^2 - \mu^2$$

$$= E(X^2) - [E(X)]^2$$

The standard deviation (SD) of a random variable x is denoted by s, then

$$\text{SD}\,(X) = \sqrt{V(X)} = \sqrt{E(X)^2 - (E(X))^2}$$

SOLVED EXAMPLES

Example 1.41. (*i*) A pair of two coins is tossed, what is the expected value?
(*ii*) A pair of dice is thrown together, find the expected value.

Solution. (*i*) Expected value or mean value $= E(X) = \mu$

$$= \sum_{i=1}^{n} p_i x_i$$ (Here X is a discrete random variable)

In tossing of two coins, probability distribution is represented in tabular form as follows :

X	0	1	2
$P(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$$\therefore \quad E(X) = \frac{1}{4}\times 0 + \frac{1}{2}\times 1 + \frac{1}{4}\times 2 = 1$$

As the probability of getting no head, one head and two heads is respectively $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$.

(*i*) In a throw of pair of dice the sum (X) is a discrete random variable which is an integer between 2 and 12 with the probabilities as given below:

X	2	3	4	5	6	7	8	9	10	11	12
$P(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$$\therefore \quad \text{Expected value} = E(X) = \mu = \frac{1}{36}\cdot 2 + \frac{2}{36}\cdot 3 + \frac{3}{36}.4 + ... + \frac{3}{36}\cdot 10 + \frac{2}{36}.11 + \frac{1}{36}\cdot 12$$

$$= \frac{252}{36} = 7$$

Note. The variance in each of the above cases is given by

$$\sigma^2 = \sum_{i=1}^{n} p_i x_i^2 - \mu^2 = \sum px^2 - \mu^2$$

In the tossing of two coins, we have

$$\Sigma p.x^2 = \frac{1}{4}.(0)^2 + \frac{1}{2}(1)^2 + \frac{1}{4}.(2)^2 = \frac{3}{2}$$

$$\therefore \quad \text{Variance} = \sigma^2 = \Sigma p.x^2 - \mu^2 = \frac{3}{2} - 1 = \frac{1}{2}$$

In case of a pair of dice, we have

$$\Sigma p.x^2 = \frac{1}{36}\cdot 4 + \frac{2}{36}\cdot 9 + \frac{3}{36}\cdot 16 + \frac{4}{36}\cdot 25 + \frac{5}{36}\cdot 36 + \frac{6}{36}\cdot 49$$

$$\frac{5}{36}\cdot 64 + \frac{4}{36}\cdot 81 + \frac{3}{36}\cdot 100 + \frac{2}{36}\cdot 121 + \frac{1}{36}\cdot 144$$

$$= \frac{1}{36}(4 + 18 + 48 + 100 + 180 + 294 + 320 + 324 + 300 + 242 + 144)$$

$$= \frac{1}{36}(1974) = \frac{329}{6}$$

$$\therefore \quad \text{Variance} = \sigma^2 = \frac{329}{6} - (7)^2 = \frac{35}{6}$$

$$\text{Standard deviation} = \sigma = \sqrt{\frac{35}{6}}$$

Example 1.42. A random variable X has the following distribution

X	–2	–1	0	1	2	3
$P(X)$	0.1	k	0.2	$2k$	0.3	k

Determine: (*i*) k, (*ii*) Mean, (*iii*) Variance.

Solution. (*i*) Since $\Sigma P(x) = 1$

$\Rightarrow \quad 0.1 + k + 0.2 + 2k + 0.3 + k = 1$

$\Rightarrow \quad 0.6 + 4k = 1$

$\therefore \quad k = 0.1$

Now the probability distribution of the random variable X is

X	–2	–1	0	1	2	3
$P(X)$	0.1	0.1	0.2	0.2	0.3	0.1

(*ii*) $\text{Mean} = \Sigma x p(x)$

$= (-2) \times (0.1) + (-1) \times (0.1) + (0) \times (0.2) + (1) \times (0.2) + (2) \times (0.3) + (3) \times (0.1)$

$= 0.8$

(*iii*) $\text{Variance} = \Sigma x^2 p(x) - [\Sigma x p(x)]^2$

$= (-2)(0.1) + (-1)(0.1) + (0)(0.2) + (1)^2 (0.2) + (2)^2 (0.3) + (3)^2 (0.1) - (0.8)^2$

$= 2.16$

Example 1.43. A bag contains 8 items of which 2 are defective. A man selects 3 items at random. Find the expected number of defective items he has drawn.

Solution. The expected number of defective items can be zero defective, one defective, two defective items. Thus, a random variable may take values 0, 1 and 2.

Now $$p_1 = P(X=0) = \frac{C(6,3) \times C(2,0)}{C(8,3)}$$

$$= \frac{6!}{3! \times 2!} \times \frac{2!}{2!} \times \frac{3! \times 5!}{8!} = \frac{20}{56}$$

$$p_2 = P(X=1) = \frac{C(6,2) \times C(2,1)}{C(8,3)}$$

$$= \frac{6!}{4! \times 1!} \times \frac{2!}{1 \times 1} \times \frac{3! \times 5!}{8!} = \frac{30}{56}$$

$$p_3 = P(X=2) = \frac{C(6,1) \times C(2,2)}{C(8,3)} = \frac{6!}{5!} \times \frac{2!}{2!} \times \frac{3! \times 5!}{8!} = \frac{6}{56}$$

Hence, the expected number of defective items drawn is

$$E(X) = p_1 x_1 + p_2 x_2 + p_3 x_3$$

$$= \frac{20}{56} \times 0 + \frac{30}{56} \times 1 + \frac{6}{56} \times 2 = \frac{42}{56} = \frac{3}{4}$$

Example 1.44. A player tossed two coins. If two heads show he wins ₹ 4. If one head shows he wins ₹ 2, but if two tails show he pays ₹ 3 as penalty. Calculate the expected value of the game to him.

Solution. Here X takes the values 0, 1, 2

Also
$$p_1 = P(X = \text{zero head}) = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}$$

$$p_2 = P(X = \text{one head}) = \frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2} = \frac{1}{2}$$

$$p_3 = P(X = \text{two heads}) = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}$$

Also when
$$X = 0 \Rightarrow x_1 = ₹\,3$$
$$X = 1 \Rightarrow x_2 = ₹\,2$$
$$X = 2 \Rightarrow x_3 = ₹\,4$$

We want to find out

$$E(x) = p_1 x_1 + p_2 x_2 + p_3 x_3 = \frac{1}{4}\times(-3)+\frac{1}{2}\times 2+\frac{1}{4}\times 4 = \frac{5}{4} = 1.25$$

Hence, $E(X) = ₹\,1.25.$

Example 1.45. Find the mean and variance of uniform probability distrubution $f(x) = \frac{1}{n}$ for $x = 1, 2 \ldots .n.$

Solution.

X	1	2	3	4	5	6	...	n
$P(X)$	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

$$\text{Mean} = E(X) = \frac{1}{n}+\frac{2}{n}+\frac{3}{n}\ldots\ldots\frac{n}{n}$$

$$= \frac{1+2+3\ldots n}{n}$$

$$= \frac{1}{n}\left[\frac{n(n+1)}{2}\right] = \frac{n+1}{2}$$

$$\text{Variance} = E(X^2) - [E(X)]^2$$

$$= \frac{1^2+2^2+\ldots n^2}{n} - \left[\frac{n+1}{2}\right]^2$$

$$= \frac{n(n+1)(2n+1)}{6n} - \left[\frac{n+1}{2}\right]^2$$

$$= \frac{n+1}{2}\left[\frac{2n+1}{3} - \frac{(n+1)}{2}\right] = \frac{n^2-1}{2}$$

Example 1.46. For discrete probability distribution.

x	0	1	2	3	4	5	6	7
f	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2+k$

Determine: (*i*) k, (*ii*) mean, (*iii*) variance, (*iv*) smallest value of x such that $P(X \le x) > \frac{1}{2}$.

Solution. (*i*) $\Sigma f(x) = 1$

$$0 + k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$$

$$10k^2 + 9k - 1 = 0$$

$$k = -1 \text{ or } \frac{1}{10}$$

$\therefore$ $k = 1$ since probability can never be negative.

x	0	1	2	3	4	5	6	7
f	0	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{100}$	$\frac{2}{100}$	$\frac{77}{100}$

(*ii*) Mean $= \Sigma x f(x)$

$$= 0 + 1\times\frac{1}{10} + 2\times\frac{2}{10} + 3\times\frac{2}{10} + 4\times\frac{3}{10} + 5\times\frac{1}{100} + 6\times\frac{2}{100} + 7\times\frac{77}{100}$$

$$= 3.66$$

(*iii*) Variance $= E(X^2) - [E(X)]^2$

$$= \left[0 + 1^2\times\frac{1}{10} + 2^2\times\frac{2}{10} + 3^2\times\frac{2}{10} + 4^2\times\frac{3}{10} + 5^5\times\frac{1}{100} + 6^2\times\frac{2}{100} + 7^2\times\frac{77}{100}\right] - (3.66)^2$$

$$= 37.7$$

(*iv*)

$$P(x \le 0) = f(0) = 0$$

$$P(x \le 1) = f(0) + f(1) = 0.1$$

$$P(x \le 2) = 0 + 0.1 + 0.2 = 0.3$$

$$P(x \le 3) = 0.3 + 0.2\ 0.5$$

$$P(x \le 4) = 0.5 + 0.3 = 0.8$$

$\therefore$ Smallest value of X such that $P(X \le x) \ge 0.5$ is 4.

Example 1.47. Two dice are thrown. Let X assign to each point (a, b) in S the maximum of its number *i.e.*, $X(a, b) = \max(a, b)$. Find the probability distribution of random variable x with $X(S) = \{1, 2, 3, 4, 5, 6\}$.

Solution. When two dice are thrown sample space S is

$$S = \begin{Bmatrix} (1,1), (1,2), (1,3), (1,4), (1,5), (1,6) \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6) \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6) \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6) \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6) \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \end{Bmatrix}$$

From the data $X[1, 1] = \max(1, 1) = 1$

$$P(X = 1) = 1/36$$

$X[(2, 1)(2, 2), (1, 2)] = 2, P(X = 2) = 3/36$

$X[(1, 3)(3, 1)(2, 3), (3, 2)(3, 3)] = 3, P(X = 3) = 5/36$

$X[(1, 4)(4, 1)(2, 4), (4, 2)(3, 4)(4, 3)(4, 4,)] = 4, P(x = 4)] = 7/36$

$X[(1, 5)(5, 1)(2, 5)(5, 2)(3, 5)(5, 3), (4, 5)(5, 4), (5, 5)] = 5, P(X = 5) = 9/36$

$X[(1, 6)(6, 2)(2, 6)(3, 6), (6, 3)(4, 6), (6, 4), (15, 6), (6, 5), (6, 6)] = 6, P(x = 6) = 11/36$

$\therefore$ Probability distribution to random variable X is

X	1	2	3	4	5	6
$P(X)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

Example 1.48. A fair coin is tossed until head or five tails occurs. Find expected number of tosses of the coins.

Solution. Probability of getting head $= 1/2 = p$

Porbability of getting tail $= 1/2 = q$

X	1	2	3	4	5	6
Outcome	H	TH	TTH	TTTH	TTTTH	TTTTT
Probability	p	qp	q^2p	q^3p	q^4p	q^5

The expected number of tosses are

$$E(x) = \Sigma x_i p_i$$

$$= 1.p + 2pq + 3.q^2p + 4qp + 5q^4p + 6q^5$$

$$= 1.\frac{1}{2} + 2\left(\frac{1}{2}\right)^2 + 3\left(\frac{1}{2}\right)^3 + 4\left(\frac{1}{2}\right)^4 + 5\left(\frac{1}{2}\right)^5 + 6\left(\frac{1}{2}\right)^5$$

$$= 1.9687 \simeq 2$$

Therefore expected number of tosses are 2.

Example 1.49. Two cards are drawn successively with replacement from a well shuffled pack of 52 cards. Find the mean and variance of the number of kings.

Solution. Let X be the random variable. The

$X =$ Number of kings obtained in two draws

Clearly, X can assume the value 0, 1, 2

$$P(\text{drawing a king}) = \frac{4}{52} = \frac{1}{13}$$

$$P(\text{not drawing a king}) = 1 - \frac{1}{13} = \frac{12}{13}$$

$$P(X = 0) = P(\text{not a king in the 1st draw and not a king in 2nd draw})$$

$$= \left(\frac{12}{13} \times \frac{12}{13}\right) = \frac{144}{169}$$

$P(X=1) = P$ (a king in the 1st drawn and not a king in the 2nd draw)

or P (not a king in the 1st draw and a king in the 2nd drawn)

$$= \left(\frac{1}{13}\times\frac{12}{13}+\frac{12}{13}\times\frac{1}{13}\right) = \frac{24}{169}$$

$P(X=2) = P$ (a king in the 1st draw and a king in the 2nd draw)

$$= \left(\frac{1}{13}\times\frac{1}{13}\right) = \frac{1}{169}$$

Hence, the probability distribution is given by

$X = x_i$	0	1	2
p_i	$\frac{144}{169}$	$\frac{24}{169}$	$\frac{1}{169}$

$\therefore$ Mean, $\mu = \Sigma p_i x_i = \left(\frac{144}{169}\times 0\right)+\left(\frac{24}{169}\times 1\right)+\left(\frac{1}{169}\times 2\right) = \frac{2}{13}$

Variance, $\sigma^2 = \Sigma p_i x_i^2 - \mu^2$

$$= \left[\left(\frac{144}{169}\times 0\right)+\left(\frac{24}{169}\times 1\right)+\left(\frac{1}{169}\times 4\right)-\frac{4}{169}\right] = \frac{24}{169}$$

Example 1.50. An industrial salesman wants to know the average number of units the sells per sales call. He checks his past sales records and comes up with the following probabilities?

Sales in units	**0**	**1**	**2**	**3**	**4**	**5**
Probability	**0.15**	**0.20**	**0.10**	**0.05**	**0.30**	**0.20**

What is the average number of units he sells per sales call?

Solution. The sales wants to known the average number of units be sells per sales call. This is the same thing as saying that the wants to know the expected value of each sales call, where a sales call is the random variable X. The expected value is calculated by the formula :

$$E(X) = p_1 x_1 + p_2 x_2 + p_3 x_3 + ...$$
$$= 0.15 \times 0 + 0.20 \times 1 + 0.10 \times 2 + 0.05 \times 3 + 0.30 \times 4 + 0.20 \times 5$$
$$= 0 + 0.2 + 0.2 + 0.15 + 1.0 = 2.75$$

Thus he would expect to sell 2.75 units on each sales call.

Example 1.51. A die is tossed twice. A 'success' is getting an odd number on a random toss. Find the variance of the number of successes.

Solution. $P(\text{success}) = \frac{3}{6} = \frac{1}{2}$

and $P(\text{failure}) = 1-\frac{1}{2} = \frac{1}{2}$

$\therefore$ $$P(X=0) = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}$$

$$P(X=1) = \frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2} = \frac{1}{2}$$

$$P(X=2) = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}$$

Hence, the probability distribution is given by

X	0	1	2
$P(X)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$\therefore$ Mean, $\mu = \Sigma p_i x_i = \frac{1}{4}\times 0+\frac{1}{2}\times 1+\frac{1}{4}\times 2 = 1$

Variance, $\sigma^2 = \Sigma p_i x_i^2 - \mu^2 = \left[\left(\frac{1}{4}\times 0\right)+\left(\frac{1}{2}\times 1\right)+\left(\frac{1}{4}\times 4\right)-1^2\right]$

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

Example 1.52. Find the expected value of the absolute difference of upturned faces in the experiment of tossing of two dies.

Solution.

Abs. DIff.	Total no. of cases	
x		$p(x)$
0	(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)	6/36
1	(1, 2) (2, 3), (3, 4), (4, 5) (5, 6)	
	(2, 1), (3, 2) (4, 3), (5, 4), (6, 5)	10/36
2	(1, 3), (2, 4), (3, 5), (4, 6), (3, 1), (4, 2) (5, 3), (6, 4)	8/36
3	(1, 4), (2, 5), (3, 6), (4, 1), (5, 2) (6, 3)	6/36
4	(1, 5), (2, 6), (5, 1) (6, 2)	4/36
5	(1, 6) (6, 1)	2/36

Hence, $$E(x) = \sum_{x=0}^{5} xp(x)$$

$$= 0\times\frac{6}{36}+1\times\frac{10}{36}+2\times\frac{8}{36}+3\times\frac{6}{36}+4\times\frac{4}{36}+5\times\frac{2}{36}$$

$$= \frac{70}{36} = \frac{35}{18}$$

Example 1.53. If X and Y are discrete random variables and K is a constant then prove that: (*i*) $E(X+K) = E(X) + K$ and (*ii*) $E(X+Y) = E(X) + E(Y)$.

Solution. Given X and Y are discrete random variable and K is a constant.

Since $$E(X) = \sum_{i=1}^{n} x_i p_i(x_i) \text{ and } \sum_{i=1}^{n} p_i = 1$$

Now
$$E(X+K) = \sum_{i=1}^{n}(X+K)p_i$$
$$= \sum_{i=1}^{n} x_i p_i + K\sum_{i=1}^{n} p_i = E(X) + K \cdot 1 \qquad \because \sum_{i=1}^{n} p_i = 1$$
$$= E(X) + K$$
$$\therefore \quad E(X+K) = E(X) + K$$

$$(X+Y) = \sum_{i=1}^{n}(x+y)p_i = \sum_{i=1}^{n} x_i p_i + \sum_{i=1}^{n} y_i p_i$$
$$= E(X) + E(Y)$$
$$E(X+Y) = E(X) + E(Y)$$

EXERCISE 1.3

1. A coin in tossed four time. If X is the number of heads observed, find the probability distribution of X.

Ans.

0	1	2	3	4
$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

2. An urn contains 5 red and 2 black balls. Two balls are randomly selected. Let X represents the number of black balls. What are the possible values of X? Is X a random variable?

Ans.

$\frac{10}{21}$	$\frac{10}{21}$	$\frac{1}{21}$

3. A random variable x has the following probability distribution.

x	0	1	2	3	4	5	6	7
$P(x)$	a	$4a$	$3a$	$7a$	$8a$	$10a$	$6a$	$9a$

(*i*) Determine the value of a.

(*ii*) Find $P(X < 3)$, $P(X \geq 4)$, $P(0 < X < 5)$

(*iii*) Give the smallest value of m for which $P(X \leq m) \leq 0.6$. **[Ans. 5]**

4. A person plays a game of tossing a coin thrice. For each head he gets ₹ 2 from the organiser, and for each tail he has to give ₹ 1.50 to the organiser. Let X denote the amount gained or lost by the person. Show that X is a random variable and exhibit it as a function on the sample space.

5. Find in probability distribution to the number of doubles in three throws of a pair of dice.

Ans.

0	1	2	3
$\frac{125}{216}$	$\frac{75}{216}$	$\frac{15}{216}$	$\frac{1}{216}$

6. Find the probability distribution of the number of successes in two tosses of a dice when a success is defined as getting a value 5 or 6.

$$\left[\textbf{Ans.}\ \begin{matrix} & 0 & 1 & 2 \\ & \frac{4}{9} & \frac{4}{9} & \frac{1}{9} \end{matrix}\right]$$

7. Two cards are drawn successively with replacement from a well-shuffled deck of 52 cards. Find the probability disribution of the number of jacks.

$$\left[\textbf{Ans.}\ \begin{matrix} & 0 & 1 & 2 \\ & \frac{144}{169} & \frac{24}{169} & \frac{1}{169} \end{matrix}\right]$$

8. A random variables X has the following probability distribution where k is some number.

$$P(x) = \begin{cases} k \text{ if } x = 0 \\ 2k \text{ if } x = 1 \\ 3k \text{ if } x = 2 \\ 0 \text{ otherwise} \end{cases}$$

(*a*) Determine the value of k. $\left[\textbf{Ans.}\ k = \frac{1}{6}\right]$

(*b*) Find $P(x < 2)$, $P(x \le 2)$, $P(x \ge 2)$. $\left[\textbf{Ans.}\ \frac{1}{2}, 1, \frac{1}{2}\right]$

9. A box contains 13 bulbs out of which 5 bulbs are defective. 3 bulbs are drawn one by one from the box without replacement. Find the probability distribution of the number of defective bulbs drawn.

$$\left[\textbf{Ans.}\ \begin{matrix} & 0 & 1 & 2 & 3 \\ & \frac{28}{143} & \frac{70}{143} & \frac{40}{143} & \frac{5}{143} \end{matrix}\right]$$

10. Two bad eggs are accidently mixed with ten good ones. Three eggs are drawn at random with replacement from this lot. Compute μ and σ^2 for the number of bad eggs drawn. $\left[\textbf{Ans.}\ \frac{1}{2}, \frac{5}{12}\right]$

11. A die is tossed thrice. A success is 'getting 1 or 6' on toss. Find the mean and the variance of the number of successes. $\left[\textbf{Ans.}\ \mu = 1 : \sigma^2 = \frac{2}{3}\right]$

12. Compute the variance of the probability distribution of the number of doublets in four throws of a pair of dice.

13. A random variable X has the following distribution.

x	1	2	3	4	8	9
$P(x)$	k	$3k$	$5k$	$7k$	$9k$	$11k$

Determine: (*i*) k, (*ii*) mean, (*iii*) $P(X) \ge 3$. $\left[\textbf{Ans.}\ (i)\frac{1}{36}, (ii) 6.14, (iii) 0.89\right]$

14. If 3 cars are drawn from a lot of 6 cars containing 2 defective cars, find :

(*i*) the probability distribution of the number of defective cars

(*ii*) the expected number of defective cars.

[**Ans.** (*i*)

X	0	1	2
$P(X)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$

(*ii*)1]

15. A random variable X has the following distribution

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

Find: (*i*) $P(X < 4)$, $P(X \geq 5)$, $P(3 < X \leq 6)$ and (*ii*) What will be the minimum value of X so that $P(X \geq 2) > 0.3$. [**Ans.** (*i*) 0.3265, 0.4898, 0.6734; (*iii*) 3]

16. A player tosses two fair coins. He wins ₹ 100/- if head appears. ₹ 200/- if two heads appear. On the other hand he loses ₹ 500/- if no head appears. Determine the expected value E of the game and is the game favourable to the player? [**Ans.** 25; Not favourable to the player]

1.12 INDEPENDENT RANDOM VARIABLE

Two random variable are said to be independent if the probability of either variable taking a particular value does not depend on the value taken by the other variable.

Theorem: The expectation of the product of two independent random variable is equal to product of their expectation.

i.e., $E(xy) = E(x)\,E(y)$ where x and y are independent random variables.

1.13. COVARIANCE

If x and y are two random variables with their respective means $\bar{x}$ and $\bar{y}$, the covariance between x and y is defined as $\text{Cov}(x, y) = E[(x - \bar{x})(y - \bar{y})]$.

Thus the expected value of product of the derivations of the two variables from their means is called their covariance.

Cor. The covariance of two independent variables is equal to zero.

Proof: If x and y are two random variables then

$$\text{Cov}(x, y) = E[(x - \bar{x})(y - \bar{y})]$$

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

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

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

[$\because$ x and y are independent random variable]

$$= \bar{x}\,\bar{y} - \bar{x}\,\bar{y} \qquad [\because\ E(x) = \bar{x},\ E(y) = \bar{y}]$$

$$= 0$$

1.14. VARIANCE OF A LINEAR COMBINATION OF RANDOM VARIABLES

If $x_1, x_2, \ldots, x_n$ be n random variables with variances $\sigma_1^2, \sigma_2^2 \ldots, \sigma_n^2$ then the variance of $u = a_1x_1 + a_2x_2 + \ldots + a_nx_n$ is given by

$$\text{Var}(u) = a_1^2 \text{Var}(x_1) + a_2^2 \text{Var}(x_2) + \ldots + a_n^2 \text{Var}(x_n) + 2a_1a_2 \text{Cov}(x_1, x_2) + \ldots + 2a_{n-1}a_n \text{Cov}(x_{n-1}, x_n)$$

Proof: We have

$$\begin{aligned} E(u) &= E(a_1x_1 + a_2x_2 \ldots + a_nx_n) \\ &= E(a_1x_1) + E(a_2x_2) \ldots + E(a_nx_n) \\ &= a_1E(x_1) + a_2E(x_2) \ldots + a_nE(x_n) \qquad [\because E(ax) = aE(x)] \end{aligned}$$

Now,

$$\mu - E(u) = a_1(x_1 - E(x_1)) + a_2(x_2 - E(x_2)) \ldots + a_n(x_n - E(x_n)]$$

Squaring both sides.

$$[\mu - E(u)]^2 = a_1^2[x_1 - E(x_1)]^2 + a_2^2[x_2 - E(x_2)]^2 + \ldots + a_n^2[x_n - E(x_n)]^2 + 2a_1a_2[x_1 - E(x_1)][x_2 - E(x_2)] + \ldots$$

Taking expectation both sides

$$\text{Var}(u) = a_1^2 \text{Var}(x_1) + a_2^2 \text{Var}(x_2) \ldots + a_n^2 \text{Var}(x_n) + 2a_1a_2 \text{Cov}(x_1, x_2) + \ldots + 2a_{n-1}a_n \text{Cov}(x_{n-1}, x_n)$$

Cor. 1. If $a_1 = a_2 = 1$ and $a_3 = a_4 \ldots = a_n = 0$

$$\text{Var}(x_1 + x_2) = \text{Var}(x_1) + \text{Var}(x_2) + 2\text{cov}(x_1, x_2)$$

Cor. 2. If $a_1 = 1$, $a_2 = -1$ and $a_3 = a_4 \ldots, a_n = 0$

$$\text{Var}(x_1 - x_2) = \text{Var}(x_1) + \text{Var}(x_2) - 2\text{ Cov}(x_1, x_2)$$

Cor. 3. If x_1 and x_2 are independent random variable then $\text{Var}(x_i \pm x_2) = \text{Var}(x_1) + \text{Var}(x_2)$

Cor. 4. If $x_1, x_2 \ldots, x_n$ are independent random variables then

$$\text{Var}(u) = a_1^2 \text{Var}(x_1) + a_2^2 \text{Var}(x_2) \ldots + a_n^2 \text{Var}(x_n)$$

Correlation coefficient

The correlation coefficient of two random variable x and y is given by

$$\rho = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)}\sqrt{\text{Var}(y)}}$$

1.15. PROPERTIES OF COVARIANCE

1. $\text{Cov}(X, X) = \text{Var}\, X$
2. If X and Y are two independent then $\text{Cov}(X, Y) = 0$
3. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
4. $\text{Cov}(aX, Y) = a\,\text{Cov}(X, Y)$
5. $\text{Cov}(X + C, Y) = \text{Cov}(X, Y)$
6. $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$

1.16. PROPERTIES OF CORRELATION

1. $-1 \le \rho(X, Y) \le 1$
2. If $\rho(X, Y) = 1$ then $Y = aX + b$ where $a > 0$
3. If $\rho(X, Y) = -1$ then $Y = aX + b$ where $a < 0$
4. $\rho(aX + b, cY + d) = \rho(X, Y)$ for $a, c > 0$

1.17. BINOMIAL DISTRIBUTION

Binomial distribution is a discrete probability distribution which is obtained when the probability p of the happening of an event is same in all the trials, and there are only two events in each trial. For example, the probability of getting a head, when a coin is tossed a number of times, must remain same in each toss, *i.e.*, $\frac{1}{2}$.

Let an experiment consisting of n trials be performed and let the occurrence of an event in any trial be called a success and its non-occurrence a failure. Let p be the probability of success and q be the probability of the failure in a single trial, where $q = 1\ p$, so that $p + q = 1$.

Let us assume that trials are independent and the probability of success is same in each trial. Let us claim that we have n trials, then the probability of happening of an event r times and failing $(n - r)$ times in any specified order is $p^r q^{n-r}$ (by the theorem on multiplication of probability). But the total number of ways in which the event can happen r times exactly in n trails is $C(n, r)$. These $C(n, r)$ ways are equally likely, mutually exclusive and exhaustive.

Therefore, the probability of r successes and $(n - r)$ failures in n trails in any order, whatever is, $C(n, r)\ p^r\ q^{n-r}$

It can also be expressed in the form

$$P(X = r) = P(r) = C(n, r)\ p^r\ q^{n-r};\ r = 0, 1, 2, 3, 4, ...n,$$

where $P(X = r)$ or $P(r)$ is the probability distribution a random variable X of the number of succeses, Giving different values of r, *i.e.*, putting $r = 1, 2, 3, ...n$, we get the corresponding probabilities ${}^nC_0\ q^n$, $C_1 q^{n-1} p$, ${}^nC_2\ q^{n-2}\ p^2$, ${}^nC_3\ q^{n-3}\ p^3$, ... p^n, which are the different terms in the Binomial expansion of $(q + p)^2$.

As a result of it, the distribution $\boldsymbol{P(r) = C(n, r)\ p^r\ q^{n-r}}$ is called Binomial probability distribution. The two independent constants. viz., n and p in the distribution are called the parameter of the distribution.

Again if the experiment (each consisting of n trials) be repeated N times, the frequency function of the Binomial distributions is given by

$$f(r) = NP(r) = NC(n, r)\ p^r\ q^{n-r}$$

The expected frequencies of 0, 1, 2, 3 ... n successes in the above set of experiment are the successive terms in the Binomial expansion of $N(q + p)^2$; where $p + q = 1$ which is also called the Binomial frequency distribution.

1.18. PROPERTIES OF A BINOMIAL DISTRIBUTION

1. It is a discrete distribution which gives the theoretical probabilities.
2. It depends on the parameters p or q, the probability of success or failure and n (the number of trails). The parameter n is always a positive integer.
3. The distrbution will be symmetrical if $p = q$.
4. The statistics of the Binomial distribution one mean = np, variance = npq; and standard deviation = $\sqrt{npq}$.

5. The mode of the binomial distribution is equal to that value X which has the largest frequency.
6. The shape and location of a binomial distribution changes as p changes for a given n or n changes for a given p.

1.19. MEAN OF BINOMIAL DISTRIBUTION

For a binomial distribution the probability function is

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

The discrete probability distribtion for the binomial distribution can be displayed as follows:

X	0	1	2	...	r	...	n
$P(X)$	${}^nC_0 q^n$	${}^nC_1 p q^{n-1}$	${}^nC_2 p^2 q^{n-2}$	...	${}^nC_r p^r q^{n-r}$	...	${}^nCp^n$

$\therefore$ Mean $(\mu) = E(X) = \sum_{r-0}^{n} rP(X = r)$

$$= {}^nC_0\, q^n \times 0 + {}^nC_1\, pq^{n-1} \times 1 + {}^nC_2\, p^2\, q^{n-2} \times 2 + \ldots + {}^nC_r\, pr\, q^{n-r} \times + \ldots + {}^nC_n\, p^n \times n$$

$$= 0 + npq^{n-1} + \frac{n(n-1)}{2!} p^2 q^{n-2} \times 2 + \ldots + np^n$$

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

$$= np\,(q+p)^{n-1} = np \quad (\because q + n)$$

$\therefore$ **Mean = np**

1.20. VARIANCE OF BINOMIAL DISTRIBUTION

Since Variance $= \Sigma px^2 - \mu^2$

Now $\Sigma px^2 = {}^nC_0\, q^n \times (0)^2 + {}^nC_1\, pq^{n-1} \times (1)^2 + {}^nC_2\, p^2\, q^{n-2} \times (2)^2$

$$+ {}^nC_3\, p^3\, q^{n-3} \times (3)^2 + \ldots + {}^nC_n\, p^n \times n^2$$

$$= 0 + n.pq^{n-1} + \frac{n(n-1)}{2!} p^2 q^{n-2} \times 4 + \frac{n(n-1)(n-2)}{3!} p^3 q^{n-3} \times 9 + \ldots + p^n \times n^2$$

Breaking second, third and following terms into parts, we get

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

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

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

$$= np + n\,(n-1)\,p^2 = np\,[1 + (n-1)\,p] = np\,[q + np]\; npq + n^2 p^2$$

$\therefore$ Variance $= npq + n^2p^2 - (np)^2 = npq.$

1.21. MODE OF BINOMIAL DISTRIBUTION

Mode is the value of r at which $p(r)$ has maximum value. Let x be the mode of binomial distribution, then

$p(r) \geq p(r+1)$ and $p(r) \geq p(r-1)$

Consider $p(r) \geq p(r+1) \Rightarrow \frac{p(r)}{p(r+1)} \geq 1$

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

$\Rightarrow (r+1)q \geq (n-r)p \Rightarrow (p+q)r \geq np - q$

$\Rightarrow r \geq np - 1 + p \Rightarrow \{(n+1)p - 1\} \leq r$ ...(*i*)

Again consider

$p(r) \geq p(r-1) \Rightarrow \frac{p(r)}{p(r-1)} \geq 1$

$\Rightarrow \frac{{}^nC_r p^r q^{n-r}}{{}^nC_{r-1} p^{r-1} q^{n-r+1}} \geq 1 \Rightarrow \frac{n-r+1}{r} \frac{p}{q} \geq 1$

$\Rightarrow (n-r+1)p \geq qr \Rightarrow r \leq (n+1)p$ ...(*ii*)

From equations (*i*) and (*ii*)

$$\{(n+1)p - 1\} \leq r \leq (n+1)p$$

Case (*i*) : If $(n+1)p$ is not an integer, then mode is the integral part of $(n+1)p$. In this case the distribution is called 'unimodal'.

Case (*ii*) : If $(n+1)p$ is an integer then both $(n+1)p$ and $\{(n+1)p - 1$ will represent modes. In this case the distribution is called 'bimodal'.

Constants of Binomial Distribution

Mean $= np$

Standard deviation $= \sqrt{npq}$

First moment or $\mu_1 = 0$

Second moment or $\mu_2 = npq$

Third moment or $\mu_3 = npq\,(q-p)$

Fourth moment or $\mu_4 = 3n^2p^2 + npq\,(1-6pq)$

$$\beta_1 = \frac{(q-p)^2}{npq}$$

$$\beta_2 = 3 + \frac{1-6pq}{npq}$$

1.22. CONDITIONS FOR APPLICATION OF BINOMIAL DISTRIBUTION

1. The variable should be discrete *i.e.*, defectives should could be 1, 2, 3, 4 or 5 etc., and never 1.5, 2.1 or 3.41 etc.
2. A dichotomy exists. In other words, the happening of events must be of two alternative. It must be either a success or failure.
3. The number of trials n should be finite and small.
4. The trials or events must be independent. The happening of one event must not affective happening of other events. In other words, statistical independence must exist.
5. The trial or events must be repeated under identical conditions.

1.23. RECURSION FORMULA OR RECURRENCE RELATION FOR BINOMIAL DISTRIBUTION

We known that for the Binomial distribution

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

and $$P(X=r+1) = {}^nC_{r+1} p^{r+1} q^{n-r-1}$$

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

$$= \frac{n!}{(r+1)!(n-r-1)!} \times \frac{r!(n-r)!}{n!} \times \frac{p^{r+1} q^{n-r-1}}{p^r q^{n-r}} = \frac{n-r}{r+1} \cdot \frac{p}{q}$$

$$\Rightarrow \quad P(X=r+1) = \frac{n-r}{r+1} \cdot \frac{p}{q} P(X=r); \; r = 1, 2, 3 \ldots..$$

which is the required recurrence formula. Applying this formula successively, we can find $P(X=1)$, $P(X=2)$, $P(X=3)$ .... if $P(X=0)$ is known.

SOLVED EXAMPLES

Example 1.54. Ten coins are thrown simultaneously. Find the probability of getting at least seven heads.

Solution. When one coin is thrown,

The probability of getting a head $= \frac{1}{2}$

$\therefore \quad p = \frac{1}{2}$

The probability of not getting a head $= 1 - \frac{1}{2} = \frac{1}{2}$

$\therefore \quad q = \frac{1}{2}$

Then P (at least 7 heads) = P (7 heads) + P(8 heads) + P(9 heads) + P(10 heads)

$$= {}^{10}C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3 + {}^{10}C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 + {}^{10}C_9 \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right) + {}^{10}C_{10} \left(\frac{1}{2}\right)^{10}$$

$$= \frac{1}{2^{10}}\left[{}^{10}C_7 + {}^{10}C_8 + {}^{10}C_9 + {}^{10}C_{10}\right]$$

$$= \frac{120+45+10+1}{1024} = \frac{176}{1024} = \frac{11}{64}$$

Example 1.55. In a lot of 200 ariticles 10 are defective, find the probability of : (*i*) no defective article, (*ii*) one defective article, (*iii*) at least one defective article, in a random sample of 20 articles.

Solution. The probability of defective article is $\frac{10}{200} = \frac{1}{20}$

$$\therefore \qquad p = \frac{1}{20}$$

The probability of non-defective article $= 1 - \frac{1}{20} = \frac{19}{20} \Rightarrow q = \frac{19}{20}$

(*i*) The probability of no defective article out of 20

$$= {}^{20}C_0 (p)^0 q^{20} = \left(\frac{19}{20}\right)^{20} \qquad \left[\because {}^{20}C_0 = 1\right]$$

(*ii*) The probability of exactly one defective article

$$= {}^{20}C_1 (p)^1 (q)^{19} = 20 \times \frac{1}{20} \times \left(\frac{19}{20}\right)^{19} = \left(\frac{19}{20}\right)^{19}$$

(*iii*) The probability of at least one will be defective

$$= 1 - [\text{probability that none will be defective}]$$

$$= 1 - {}^{20}C_{20}\left(\frac{19}{20}\right)^{20} = 1 - \left(\frac{19}{20}\right)^{20}$$

Example 1.56. If on an average, one ship out of 10 is wrecked, find the probability that out of 5 ships expected to arrive the port, at least four will arrive safely.

Solution. p be the probability of a ship arriving safely $= 1 - \frac{1}{10} = \frac{9}{10}$

$$q = 1 - \frac{9}{10} = \frac{1}{10}$$

Binomial distribution is $\left(\frac{1}{10} + \frac{9}{10}\right)^5$

Probability that at least four ships out of five arrive safely

$$= P(4) + P(5) = {}^5C_4\left(\frac{1}{10}\right)^1\left(\frac{9}{10}\right)^4 + {}^5C_5\left(\frac{1}{10}\right)^0\left(\frac{9}{10}\right)^5$$

$$= \left(\frac{9}{10}\right)^4 \frac{14}{10} = \left(\frac{9}{10}\right)^4 \frac{7}{5} = 0.91854$$

Example 1.57. The probability that a man aged 60 will live to be 70 is 0.65. What is the probability that out of 10 men aged 60 now, at least 7 would live to be 70?

Solution. Probability of survival upto the age of 70

$$= p = 0.65$$

Probability of non-survival upto the age of 70

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

Probability that out of 10 such men at least 7 would survive as desired

= Probability that exactly 7 would survive +
Probability that exactly 8 would survive +
Probability that exactly 9 would survive +
Probability that exactly 10 would survive

$$= P(7) + P(8) + P(9) + P(10)$$
$$= {}^{10}C_7 p^7 q^3 + {}^{10}C_8 p^8 q^2 + {}^{10}C_9 p^9 q + {}^{10}C_{10} p^{10}$$
$$= 120\, p^7 q^3 + 45\, p^8 q^2 + 10\, p^9 q + p^{10}$$
$$= p^7 (120\, q^3 + 45\, pq^2 + 10\, p^2 q + p^3)$$
$$= (0.65)^7 [120 \times (0.35)^3 + 45\,(0.65)\,(0.35)^2 + 10\,(0.65)^2\,(0.35) + (0.65)^3]$$
$$= 0.514, \text{ the required result.}$$

Example 1.58. Six dice are thrown together at a time, the process is repeated 729 times. How many times do you expect at least three dice to have 4 to 6?

Solution. The chance of getting 4 or 6 with one dice is

$$\frac{2}{6} \text{ i.e., } p = \frac{1}{3} \quad \text{and} \quad q = 1 - \frac{1}{3} = \frac{2}{3}$$

In one throw of six dice together, we have probability of getting at least 3 dice to have 4 or 6.

$$= P(3) + P(4) + P(5) + P(6)$$
$$= {}^6C_3 p^3 q^3 + {}^6C_4 p^4 q^2 + {}^6C_5 p^5 q + {}^6C_6 p^6$$
$$= 20\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^3 + 15\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^2 + 6\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right) + \left(\frac{1}{6}\right)^6$$
$$= \frac{1}{(3)^6}[160 + 60 + 12 + 1] = \frac{233}{(3)^6}$$

Now the process is repeated 729 times

$\therefore$ Required number of times at least 3 dice have 4 or 6

$$= 729 \times \frac{233}{(3)^6} = 233, \text{ the required result.}$$

Note. In the above case the binomial distribution is $N(q + p)^n$ where $N = 729$, $n = 6$

Example 1.59. If the sum of the mean and the variance of binomial distribution of 5 trials is 4.8, find the distribution.

Solution. Let the required binomial distribution be ${}^nC_r\, p^r\, q^{n-r}$ where n = number of trials = 5

Mean of the distribution = np

and the variance of the distribution = npq

By the given condition

$$np + npq = 4.8$$

$$\Rightarrow \quad 5p + 5pq = 4.8$$

$$5p(1+q) = 4.8 \qquad [\because p = 1-q]$$

$$\Rightarrow \quad 50(1-q^2) = 48 \quad \Rightarrow \quad 50 - 50q^2 = 48$$

$$\Rightarrow \quad 50q^2 = 2 \Rightarrow q = \frac{1}{5}$$

$$\therefore \quad p = 1-q = 1-\frac{1}{5} = \frac{4}{5}$$

Hence, the required binomial distribution is ${}^5C_r\left(\frac{4}{5}\right)^r\left(\frac{1}{5}\right)^{5-r}$

Example 1.60. The probability that a bomb dropped from a place will strike the target is $\frac{1}{5}$. If six boms are dropped, find the probability that : (*i*) exactly two will strike the target, (*ii*) at least two will strike the target.

Solution. The probabilities of 0,1, 2 ..., successes are given by the respective terms in the expansion of

$$(q+p)^n = \left(\frac{4}{5}+\frac{1}{5}\right)^6, \text{ since } p = \frac{1}{5}, q = \frac{4}{5} \text{ and } n = 6.$$

$\therefore P(2)$ = The probability that exactly two bombs will strike the target

$$= {}^6C_2 p^2 q^4 = \frac{6.5}{1.2}\cdot\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^4 = 0.246$$

The probability that at least 2 bombs will strike the target

$$= 1 - [P(0) + P(1)]$$

$$= 1 - q^6 - {}^6C_1\, q^5 p = 1 - (0.8)^6 - 6(0.2)\,(0.8)^5$$

$$= 1 - 0.2621 - 0.3932 = 0.345$$

Example 1.61. Assuming that half the population are consumers of rice so that the chance of an individual being a rice consumer is $\frac{1}{2}$ and assuming that 100 investigations each take 10 individuals to see whether they are rice consumers. How many investigations would you expect to report that three people or less consumers?

Solution. Here $p = \frac{1}{2}, q = \frac{1}{2}, n = 10, N = 100$

$\therefore$ The probability that r persons out of 10 persons are consumers of rice is given by

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

$\therefore$ The expected number of investigators (*i.e.*, expected frequencies) who would report that three or less people were consumers of rice.

$$= 100\,[P(0) + P(1) + P(2) + P(3)]$$

$$= 100\left[{}^{10}C_0\left(\frac{1}{2}\right)^{10} + {}^{10}C_1\left(\frac{1}{2}\right)^{10} + {}^{10}C_2\left(\frac{1}{2}\right)^{10} + {}^{10}C_3\left(\frac{1}{2}\right)^{10}\right]$$

$$= \frac{100}{2^{10}}[1 + 10 + 45 + 120] = \frac{17600}{1024} = 17 \text{ approx.}$$

Example 1.62. A die is thrown 5 times. Getting an even number greater than 2 is considered a success. Calculate $P(X = r)$ for $r = 1, 2, 3, 4, 5$ from recurrence formula.

Solution. Let p be the probability of getting an even number greater than 2 on a die.

$$\Rightarrow \quad p = \frac{2}{6} = \frac{1}{3}$$

$$\therefore \quad q = 1 - p \quad \Rightarrow \quad q = 1 - \frac{1}{3} = \frac{2}{3}$$

$$\therefore \quad \frac{p}{q} = \frac{1}{2}. \quad \text{Also} \quad n = 5$$

$P(X) = 0 =$ Probability of no success in 5 trials

$$= {}^5C_0(q)^5 = \left(\frac{2}{3}\right)^5 = 0.1317$$

Recurrence formula for binomial distribution is

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

$$= \frac{5 - r}{r + 1} \cdot \left(\frac{1}{2}\right) P(X = r) \qquad \ldots(i)$$

Putting $r = 0$, in (i), $P(X = 1) = 5\left(\frac{1}{2}\right) P(X = 0) = 5\left(\frac{1}{2}\right)(0.1317) = 0.3292$

Putting $r = 1$, in (i), $P(X = 2) = 2\left(\frac{1}{2}\right) P(X = 1) = P(X = 1) = 0.3292$

Putting $r = 2$, in (i), $P(X = 3) = (1)\left(\frac{1}{2}\right) P(X = 2) = \frac{1}{2} \cdot (0.3292) = 0.1646$

Putting $r = 3$, in (i), $P(X = 4) = \frac{2}{4} \cdot \frac{1}{2} P(X = 3) = \frac{1}{4}(0.1646) = 0.0412$

Putting $r = 4$, in (i), $P(X = 5) = \frac{1}{5} \cdot \frac{1}{2} P(X = 4) = \frac{1}{10}(0.0412) = 0.0041.$

Example 1.63. Out of 800 families with 4 children each, how many families would be expected to have: (i) 2 boys and 2 girls, (ii) at least one boy, (iii) no girl, (iv) at most two girls? Assume equal probabilities for boys and girls.

Solution. Since probability for boys and girls are equal

$$p = \text{Probability of having a boy} = \frac{1}{2}$$

$$q = \text{Probability of having a girl} = \frac{1}{2}$$

$$n = 4, \ N = 800$$

The binomial distribution is $800\left(\frac{1}{2} + \frac{1}{2}\right)^4$.

(i) The expected number of families having 2 boys and 2 girls

$$= 800 \cdot {}^4C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = 800 \times 6 \times \frac{1}{16} = 300$$

(ii) The expected number of families having at least one boy

$$= 800\left[{}^4C_1\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) + {}^4C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 + {}^4C_3\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 + {}^4C_4\left(\frac{1}{2}\right)^4\right]$$

$$= 800 \times \frac{1}{16}[4 + 6 + 4 + 1] = 750$$

(iii) The expected number of families having no girl having 4 boys

$$= 800 \times {}^4C_4\left(\frac{1}{2}\right)^4 = 50$$

(iv) The expected number of families having at most two i.e., having at least 2 boys

$$= 800\left[{}^4C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 + {}^4C_3\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 + {}^4C_4\left(\frac{1}{2}\right)^4\right]$$

$$= 800 \times \frac{1}{16}[6 + 4 + 1] = 550.$$

Example 1.64. A student obtained the following answer to a certain problem given to him. Mean = 2.4; variance = 3.2 for a binomial distribution. Comment on the result.

Solution. The mean of binomial distribution is np and variance npq. We are given mean $= np = 2.4$

$$\text{Variance} = npq$$

$$2.4q = 3.2$$

$$q = \frac{3.2}{2.4} = 1.333$$

Since the value of q is greater than 1, the given results are inconsistent.

Example 1.65. Ten coins are tossed 1024 times and the following frequencies are observed. Compare these frequencies with the expected frequencies :

Number of heads	0	1	2	3	4	5	6	7	8	9	10
Frequencies	2	10	38	106	188	257	226	128	59	7	3

Solution. Here $n = 10, N = 1024$

$p =$ The change of getting a head in one toss $= \frac{1}{2}$

$\therefore \quad q = 1 - p = \frac{1}{2}$

The expected frequencies are the respective terms of the binomial $1024\left(\frac{1}{2}+\frac{1}{2}\right)^{10}$

The frequency of r heads $(0 \geq r \geq 10)$ is

$$= 1024.{}^{10}C_r\left(\frac{1}{2}\right)^{10-r}.\left(\frac{1}{10}\right) = 1024 \times {}^{10}C_r\left(\frac{1}{2}\right)^{10} = {}^{10}C_r$$

Hence, we have the following comparison.

Number of heads	0	1	2	3	4	5	6	7	8	9	10
Observed frequency	2	10	38	106	188	257	226	128	59	7	3
Expected frequency	1	10	45	120	210	252	210	120	45	10	1

$({}^{10}C_0, {}^{10}C_1, {}^{10}C_2$ and so on.)

Example 1.66. Probability of man hitting a target is $\frac{1}{3}$.

(*a*) If the fires of 6 times, what is the probability of hitting: (*i*) at most 5 times, (*ii*) at least 5 times, (*iii*) exactly once

(*b*) If he fires so that the probability of his hitting target atleast once is greater than $\frac{3}{4}$, find *n*.

Solution. (*a*) Given $p = 1/3, q = 1 - 1/3 = 2/3$ $n = 6$

(*i*) The probability of hitting the target almost 5 times.

$$P(X \leq 5) = 1 - P(X > 5) = 1 - P(X = 6)$$

$$= 1 - \left[{}^6C_6\left(\frac{1}{3}\right)^6\left(\frac{2}{3}\right)^0\right] = 1 - \frac{1}{729} = \frac{728}{729}$$

(*ii*) The probability of hitting the target atlesat 5 times

$$P(X \geq 5) = P(X = 5) + P(X = 6)$$

$$= {}^6C_5\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^1 + {}^6C_6\left(\frac{1}{3}\right)^6\left(\frac{2}{3}\right)^0 = \frac{13}{729}$$

(*iii*) The probability of hitting the target exactly once

$$P(X=1)\ {}^{6}C_1\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^5 = \frac{192}{729}$$

(*b*) If he fires so that the probability of his hitting the target atleast once is greater than 3/4 then

$$\Rightarrow \quad P(X \geq 1) = 3/4$$

$$\Rightarrow \quad 1 - P(X<1) > 3/4$$

$$\Rightarrow \quad 1 - P(X=0) > 3/4$$

$$\Rightarrow \quad 1 - (2/3)^n > 3/4$$

$$\Rightarrow \quad 1/4 > (2/3)^n$$

$$\Rightarrow \quad \frac{1}{2^2} > \left(\frac{2}{3}\right)^n$$

$$\Rightarrow \quad 3n > 2^{n+2}$$

This inequality is satisfied for $n = 4$

$\therefore$ He must fire 4 times to that probability of hitting the target atleast once is greater than 3/4.

Example 1.67. A student takes a true-false examination consisting of 8 questions. He guesses each answer. The guesses are made at random. Find the smallest value of *n* so that the probability of guessing atleast *n* correct answers is less than 1/2.

Solution. For given data, we have to find

$$P(X \geq n) < 1/2$$

Now probability of guessing a correct answer is 1/2 and guessing a wrong answer is 1/2

$$\therefore \quad p = 1/2, q = 1/2, n = 8$$

Using Binomial distribution.

$$P(X \geq n) < 1/2$$

$$\Rightarrow \quad 1 - P(X<n) < 1/2$$

$$\Rightarrow \quad 1 - [P(X=0) + P(X=1) \ldots P(X=n-1)] < 1/2$$

$$\Rightarrow \quad P(X=0) + P(X=1) \ldots P(X=n-1) > 1/2$$

$$\Rightarrow \quad {}^{8}C_0\left(\frac{1}{2}\right)^8 + {}^{8}C_1\left(\frac{1}{2}\right)^8 + \ldots {}^{8}C_{n-1}\left(\frac{1}{2}\right)^8 > \frac{1}{2}$$

$$\Rightarrow \quad \left(\frac{1}{2}\right)^8 \left[{}^{8}C_0 + {}^{8}C_1 + \ldots + {}^{8}C_{n-1}\right] > \frac{1}{2}$$

This inequality is satisfied it $n - 1 = 4$

$$\Rightarrow \quad n = 5$$

Example 1.68. Fit a binomial distribution to following data, when tossing 5 coins.

x	0	1	2	3	4	5
f	2	14	20	34	22	8

Solution.

$$\Sigma f = 100 = N$$

$$n = 5$$

$$\bar{x} = \text{mean}$$

$$= \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{0\times 2 + 1\times 14 \ldots + 5\times 8}{100}$$

$$= 2.84$$

$$\text{Mean} = \bar{x} = n.p = 2.84$$

$$\Rightarrow \quad p = 0.57 \text{ and } q = 0.43 \qquad [\because n = 5]$$

Using binomial distribution

$$P(X = x) = {}^nC_x p^x q^{n-x} \; ; \; x = 0, 1, 2 \ldots n$$

and expected frequency obtained from

$$f(x) = N.\, P(x)$$

x	f	$p(x) = {}^nC_x p^x q^{n-x}$	$f(x) = N.P(x)$
0	2	${}^5C_0 (0.57)^0 (0.43)^5 = (0.43)^5$	$100 \times (0.43)^5 \approx 1$
1	14	${}^5C_1 (0.57)^1 (0.43)^4 = 0.098$	$100 \times (0.098) \approx 10$
2	20	${}^5C_2 (0.57)^2 (0.43)^3 = 0.260$	$100 \times (0.260) \approx 26$
3	34	${}^5C_3 (0.57)^3 (0.43)^2 = 0.342$	$100 \times (0.342) \approx 34$
4	22	${}^5C_4 (0.57)^4 (0.43) = 0.224$	$100 \times (0.224) \approx 22$
5	8	$C_5 (0.57)^5 (0.43)^0 = 0.059$	$100 \times (0.059) \approx 6$

Example 1.69. In a Binomial distribution consisting of 5 independent trials, probability of 1 and 2 success are 0.4096 and 0.2048 respectively. Find the parameter 'p' of distribution.

Solution. Given $n = 5$

For Binomial distribution

$$P(X = x) = {}^nC_x p^x q^{n-x}; \; x = 0, 1, 2, 3 \ldots n$$

$$P(X = 1) = {}^nC_1 p^1 q^{n-1} = 0.4096 \qquad \ldots(i)$$

$$P(X - 2) = {}^nC_2 p^2 q^{n-2} = 0.2048 \qquad \ldots(ii)$$

Dividing eqn. (*ii*) with eqn. (*i*)

$$\frac{{}^5C_2 p^2 q^3}{{}^5C_1 p^1 q^4} = \frac{10p}{5q} = \frac{0.4096}{0.2048}$$

$$\Rightarrow \quad \frac{2p}{1-p} = 1/2$$

$$\Rightarrow \quad 4p = 1 - p$$

$$\Rightarrow \quad p = 1/5$$

EXERCISE 1.4

1. During war, 1 ship out of 9 was sunk on an average in making a certain voyage. What was the probability that exactly 3 out of the convey of 6 ships would arrive safely? $\left[\text{Ans.}\frac{10240}{9^6}\right]$

2. The incidence of occupational disease in an industry in such that the workers have a 20% change of offering from it. What is the probability that out of six workers chosen at random, four or more will suffer from the disease? $\left[\text{Ans.}\frac{53}{3125}\right]$

3. The probability that a pen manufactured by a company will be defective is $\frac{1}{10}$. If 12 such pens are manufactured, find the probability that : (*i*) exactly 2 will be defective, (*ii*) none will defective, (*iii*) at least two will be defective.

4. A dice is thrown. If "getting an odd number" is a "success", what is the probability of :
   (*i*) 5 successes
   (*ii*) at least 5 successes
   (*iii*) almost 5 successes [**Ans.** (*i*) $\frac{3}{32}$, (*ii*) $\frac{7}{64}$, (*iii*) $\frac{63}{64}$]

5. If on an average one ship in every ten is wrecked, find the probability that out 5 ships expected to arrive, 4 at least will arrive safely. $\left[\text{Ans. } \frac{7}{5}\left(\frac{9}{10}\right)^4\right]$

6. Five cards are drawn successively with replacement from a well-shuffled pack of 52 cards. What is the probability that:
   (*i*) all the five cards are spades
   (*ii*) only 3 cards are spades
   (*iii*) none is a spade [**Hints.** Number of spades : 13] $\left[\text{Ans. } (i)\left(\frac{1}{4}\right)^5, (ii)\,90\left(\frac{1}{4}\right)^5, (iii)\left(\frac{3}{4}\right)^5\right]$

7. State reason to justify whether the following statement is true or false. "The mean of a Binomial distribution is 6 and standard deviation is 3". [**Ans.** false]

8. *A* and *B* take turns in throwing dice, the first to throw 10 being the winner. If *A* throws firstly, show that they have change of winning as 12 : 11.

9. In a bombing action there is 50% chance that any bomb will strike target. Two direct hits are needed to destroy and target completely. How many bombs are required to be dropped to give a 99% chance of better of completely destroying the target. [**Ans.** 11]

10. Out of 800 families with 5 children each, how many would you expect to have: (*i*) 3 boys, (*ii*) 5 girls, (*iii*) either 2 or 3 boys? Assume equal probabilities for boys and girls.
    [**Ans.** (*i*) 250, (*ii*) 25, (*iii*) 500]

11. In sampling a large number of parts manufactured by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples, how many would be expected to contain at least 3 defective parts. [**Ans.** 458]

12. The probability that a bulb produced by a factory will fuse after 150 days of use is $\frac{1}{20}$. Find the probability that out of 5 such bulbs :

(*i*) None

(*ii*) not more than one

(*iii*) more than one

(*iv*) at least one, will fuse after 150 days of use.

$$\left[\textbf{Ans. } (i)\left(\frac{19}{20}\right)^5, (ii)\frac{24}{20}\left(\frac{19}{20}\right)^4, (iii)\,1-\frac{23}{20}\left(\frac{19}{20}\right)^4, (iv)\,1-\left(\frac{19}{20}\right)^5\right]$$

13. Four coins are tossed 160 times. The number of times r heads occur ($r = 0, 1, 2, 3, 4$) is given below:

r	0	1	2	3	4
No. of times	8	34	69	43	6

Fit a binomial distribution to this data on the hypothesis that coins are unbiased.

$$\left[\textbf{Ans. } \begin{matrix} r = 0 & 1 & 2 & 3 & 4 \\ f(r) = 10 & 40 & 60 & 40 & 10 \end{matrix}\right]$$

14. If successive trials are independent and the probability of success on any trial is p, shown that the first success occurs on the nth trial is

$$p(1-p)^{n-1},\ n = 1, 2, 3, \ldots.$$

15. The mean of a Binomial distribution is 3 and variance is 4. Give your comments.

16. If the chance that one of the ten telephone lines is busy at an instant is 0.2.

(*i*) What is the chance that 5 of the lines are busy?

(*ii*) What is the probability that all the lines are busy? **[Ans.** (*i*) 0.02579, (*ii*) 21.024×10^{-7}**]**

17. The following data are the number of seeds germinating out of 10 on damp filter for 80 sets of seeds. Fit a binomial distribution of this data :

x	0	1	2	3	4	5	6	7	8	9	10	Total
f	6	20	28	12	8	6	0	0	0	0	0	80

[Ans. $80(0.7825 + 0.2175)^{10}$**]**

18. Mark the correct answer.

(*a*) The probability that a man hit a target is given as $\frac{1}{5}$. Then his probability of atleast one hit in 10 shots is

(*i*) $1-\left(\frac{4}{5}\right)^{10}$ (*ii*) $\left(\frac{1}{5}\right)^{10}$ (*iii*) $1-\left(\frac{1}{5}\right)^{10}$ (*iv*) None

(*b*) 8 coins are tossed simultaneously. The probability of getting at least 6 heads is

(*i*) $\frac{57}{64}$ (*ii*) $\frac{229}{256}$ (*iii*) $\frac{7}{64}$ (*iv*) $\frac{37}{256}$

[Ans. (*a*) (*i*), (*b*) (*iv*)**]**

20. Fill in the blanks :

(*i*) The probability of getting number 5 exactly two times in five throws of an unbiased dice is...

(*ii*) If the probability of hitting a target is 5% and 5 shots are fired independently, the probability that the target will be hit at least once is .....

[**Ans.** (*a*) $10\left(\frac{5^3}{6^5}\right)$, (*b*) $1-(0.95)^5$]

1.24. POISSON DISTRIBUTION

It is due to the French Mathematician S.D. Poisson (1781–1840) who published its derivation in 1837. It is a discrete probability distribution which has the following characteristics:

(*i*) It is the limiting form of the Binomial distribution as n becomes infinitely large *i.e.*, $n \to \infty$ and p, the constant probability of success for each trial becomes indefinitely small *i.e.*, $p \to 0$ in such a manner that $np = m$ remains a finite number.

(*ii*) It consists of a single parameter m only. The entire distribution can be obtained once m is known.

It has wide applications in physical, engineering and management sciences as well as economics, operations research and reliability technology.

Poisson distribution occurs when there are events which do no occur as outcomes of definite number of trials of an experiment but which occurs at random point of time and space wherein our interest lies only in the number of occurrences of the events. Here we are concerned with processes taking place over continuous intervals of time such as arrival of a telephone calls at a switch board, or the passing by cars of an electric checking device. That is why is has important applications is queuing theory where we are interested, for example, in the number of aircraft arriving at an air field, the number of ships and trucks arriving to be unloaded at a receiving dock etc.

Some of the situations where the probability of the event remains very small and the Poisson distribution is applicable, are: (*i*) the occurrence of accidents in a factory in a given period, (*ii*) the number of defective articles produced by some factory in a fixed time period, (*iii*) the number of twins births per year in some hospital, (*iv*) and the number of deaths due to snake bite in a city per year, (*v*) the number of printing mistakes whch page of the book, (*vi*) the number of persons blind in a large city every year etc.

1.25. BINOMIAL APPROXIMATION TO POISSON DISTRIBUTION

We will derive Poisson distribution as a limiting case of Binomial distribution when $p \to 0, n \to \infty$ such that $np = m$ (a finite quantity). We know that in a Binomial distribution, the probability of r successes is given by

$$P(r) = {}^nC_r\, q^{n-r} = \frac{n(n-1)(n-2)...(n-r+1)}{r!}(1-p)^{n-r}\, p^r$$

on expanding ${}^nC_r$ and replacing q by $(1-p)$.

$$P(r) = \frac{n(n-1)(n-2)...(n-r+1)}{r!}\left(1-\frac{m}{n}\right)^{n-r}\left(\frac{m}{n}\right)^r$$

since $$np = m \text{ or } p = \frac{m}{n}$$

or
$$P(r) = \frac{n(n-1)(n-2)...(n-r+1)}{r!n^r}\left(1-\frac{m}{n}\right)^n\left(1-\frac{m}{n}\right)^{-r}\cdot m^r$$
$$= \frac{n}{n}\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)...\left(\frac{n-r+1}{n}\right)\left(1-\frac{m}{n}\right)^n\times\left(1-\frac{m}{n}\right)^{-r}\cdot\frac{m^r}{r!}$$
$$= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{r-1}{n}\right).\left(1-\frac{m}{n}\right)^n\times\left(1-\frac{m}{n}\right)^{-r}\frac{m^r}{r!}$$

Now for a given value of r, when $n \to \infty$, we have

$$\lim_{n\to\infty}\left[\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{r-1}{n}\right)\right] = 1,$$

$$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^{-r} = 1 \text{ and } \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = e^{-m}$$

As such in the limiting form

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

This is the probability of r successes for Poisson distribution. For $r = 0, 1, 2, 3, ...$, we get the probabilities of 0, 1, 2, 3, ... successes as

$$P(0) = e^{-m},\ P(1) = me^{-m},\ P(2) = \frac{m^2}{n!}e^{-m}$$

$$P(3) = \frac{m^3}{3!}e^{-m} \text{ ... and so on}$$

Note : 1. The sum of the probabilities $P(r)$ for $r = 0, 1, 2, 3, ...$ is 1.

i.e.,
$$\Sigma P(r) = P(0) + P(1) + P(2) + P(3) + ...$$
$$= e^{-m}\left(1+m+\frac{m^2}{2!}+\frac{m^3}{3!}+...\right) = e^{-m}\times e^m = 1.$$

2. Poisson distribution possesses only one parameter 'm'. If we consider the length of interval as 'd' instead of unit length, the average number of occurrences in 'd' length of interval is 'md' Thus, the probability function in this situation is,

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

1.26. CONDITIONS UNDER WHICH POISSON DISTRIBUTION IS USED

1. The random variable x should be discrete.
2. A dichotomy exists, *i.e.*, the happening of the event must be of two alternatives such as success and failure, ocurrences and non-occuraence etc.
3. It is applicable in those case where the number of trials n is very large and the probabilities of success p is very small but the mean $np = m$ is finite.

4. p should be very small (close of zero). If $p \to 0$, then the distribution is J-shaped and unimodal.
5. Statistical independence is assumed. In other words, it is applicable to those cases when the happening of an even does not affect the happening of the other events.

1.27. MEAN OF POISSON DISTRIBUTION

Poisson distribution is $P(r) = \dfrac{e^{-m}.m^r}{r!}, r = 0, 1, 2,...$

$$\text{Mean, } (\mu) = \sum_{i=0}^{i=n} x_i p_i = p_1 x_1 + p_2 x_2 + p_3 x_3 \ldots p_n x_n$$

$$= 0e^{-m} + 1e^{-m} + 2\frac{m^2e^{-m}}{2!} + 3\frac{m^3e^{-m}}{3!} + 4\frac{m^4e^{-m}}{4!} + \ldots$$

$$= me^{-m} + m^2e^{-m} + \frac{m^3e^{-m}}{2\times1} + \frac{m^4e^{-m}}{3.2.1} + \ldots$$

$$= me^{-m}\left(1 + m + \frac{m^2}{2.1} + \ldots\right)$$

$$= me^{-m}\, e^m \qquad \left(\because e^m = 1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \ldots + \right)$$

$$= me^{-m+n} = me^{0} = m$$

$\therefore$ The mean of the Poisson distribution is m.

1.28. VARIANCE OF POISSON DISTRIBUTION

$$\text{Variance } (\sigma^2) = (\Sigma x_i^2 p_1) - \mu^2$$

Now $$\sum_{i=1}^{i=n} p_i x_i^2 = 0.e^{-m} + 1\cdot me^{-m} + 2^2.\frac{m^2e^{-m}}{2!} + 3^2.\frac{m^3e^{-m}}{3!} + 4^2.\frac{m^4e^{-m}}{4!} + .$$

$$= me^{-m} + 2m^2e^{-m} + \frac{3m^3e^{-m}}{2\times1} + \frac{4m^4e^{-m}}{3\times2\times1} + \ldots$$

$$= me^{-m}\left[1 + 2m + \frac{3m^2}{2\times1} + 4\frac{m^3}{3\times2\times1} + \ldots\right]$$

$$= me^{-m}\left[\left(1 + m + \frac{m^2}{2\times1} + \frac{m^3}{3\times2\times1} + \ldots\right) + \left(m + \frac{2m^2}{2} + \frac{3m^3}{3\times2\times1} + \ldots\right)\right]$$

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

$$= me^{-m}\,(e^m + me^m) = m\,(m+1)\,e^{-m}\,e^m = m\,(m+1)$$

$\therefore$ $$\text{Variance} = \Sigma x_i^2\, p_i = \mu^2\, m\,(m+1) - m^2 = m^2 + m - m^2 = m$$

$$\text{Standard deviation} = \sqrt{\text{Variance}} = \sqrt{m}$$

1.29. MODE OF POISSON DISTRIBUTION

Mode is a value of at which $p(r)$ has maximum value, then

$$p(r) = p(r+1) \quad \text{and} \quad p(r) = p(r-1)$$

Consider $p(r) \geq p(r+1)$

i.e., $$\frac{e^{-m}}{r!} \geq \frac{e^{-m}m^{r+1}}{(r+1)!} \Rightarrow \frac{r+1}{m} \geq 1$$

$\Rightarrow$ $$r+1 \geq m \Rightarrow r \geq m-1$$

i.e., $$m-1 \leq r \qquad ...(i)$$

Similarly consider

$$p(r) \geq p(r-1) \text{ then we get } r \geq m \qquad ...(ii)$$

From eqns. (i) and (ii)

$$m-1 \leq r \leq m$$

Therefore the mode of the Poisson distribution lies between $(m-1)$ and m.

Constants of Poisson Distribution

$$\mu_1 = 0, \mu_2 = m, \mu_3 = m, \mu_4 = m + 3m^2$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{m+3m^2}{m^2} = 3 + \frac{1}{m}$$

1.30. RECURRENCE FORMULA FOR THE POISSON DISTRIBUTION

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

and $$P(r+1) = \frac{e^{-m} \cdot m^{r+1}}{(r+1)!}$$

Then $$\frac{P(r+1)}{P(r)} = \frac{\frac{e^{-m} \cdot m^{r+1}}{(r+1)!}}{\frac{e^{-m} \cdot m^r}{r!}} = \frac{e^{-m} \cdot m^{r+1}}{e^{-m} \cdot m^r} \times \frac{r!}{(r+1)!} + \frac{m}{r+1}$$

$\Rightarrow$ $$P(r+1) = \frac{m}{r+1} P(r); \; r = 0, 1, 2, \ldots$$

which is the required recurrence formula for Poisson distribution. With this formula, we can find $P(1), P(2), P(3), P(4), \ldots$ if $P(0)$ is given.

SOLVED EXAMPLES

Example 1.70. The number of telephone calls arriving on an internal switch board of an office is 90 per hour. Find the probability that at the most 1 to 3 calls in a minute on the board arrive. (Use $e^{-1.5} = 0.223$)

Solution. $m = \text{mean} = \dfrac{90}{60} = 1.5$. Obviously, X will follow Poisson distribution.

Now the probability that at the most 1 to 3 calls in one minute

$$= P(1) + P(2) + P(3)$$

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

$$P(1) = \frac{e^{-1.5}(1.5)^1}{1!} = 1.5 \times e^{-1.5} = 1.5 \times 0.223$$

$$P(2) = \frac{e^{-1.5}(1.5)^2}{2!} = 1.125 \times e^{-1.5} = 1.125 \times 0.223$$

$$P(3) = \frac{e^{-1.5}(1.5)^3}{3!} = 0.562 \times e^{-1.5} = 0.562 \times 0.223$$

$\therefore$ Required probability $= 0.233\ (1.5 + 1.125 + 0.562) = 0.711$

Example 1.71. Suppose a book of 585 pages contains 43 typographical errors. If these errors are randomly distributed throughout the book, what is the probability that 10 pages, selected at random, will be free from errors? (Use $e^{-0.735} = 0.4795$).

Solution. Here $p = \dfrac{43}{585} = 0.0735$ and $n = 10$.

$\therefore$ $$m = np = 10 \times 0.0735 = 0.735$$

Clearly, p is very small and n is large.

So, it is a case of Poisson distribution.

Let X denote the number of errors in 10 pages.

Then, $$P(X = r) = \frac{e^{-m} \cdot m^r}{r!} = \frac{e^{-0.735} \times (0.735)^r}{r!}$$

$\therefore$ $$P(\text{no error}) = P(X = 0) = \frac{e^{-0.735} \times (0.735)^0}{0!} = e^{-0.735} = 0.4795$$

Hence, the required probability is 0.4795.

Example 1.72. Average number of accidents on any day on a national highway is 1.8. Determine the probability that the number of accidents are: (*i*) at least one, (*ii*) at most one. (Given $e^{-1.8} = 0.16529$).

Solution. The probability function of the Poisson distribution is

$$P(X = r) = \frac{e^{-m} \cdot m^r}{r!} \qquad (r = 0, 1, 2, 3, \ldots)$$

Given that $m = 1.8$

(*i*) $P(X \geq 1)$: $P(X \geq 1) = 1 - p(r < 1)$

$$= 1 - p(r = 0)$$

$$= 1 - \frac{e^{-m} m^0}{0!} = 1 - \frac{e^{1.8}(1.8)^0}{0!}$$

$$= 1 - e^{-1.8} = 1 - 0.16529 = 0.8347$$

$\therefore$ The probability the number of accidents at least one is 0.8347.

(*ii*) $P(X \leq 1)$: $P(X \leq 1) = p(X = 0) + p(X = 1)$

$$= \frac{e^{-m} m^0}{0!} + \frac{e^{-m} m^1}{1!} = \frac{e^{-1.8}(1.8)^0}{0!} + \frac{e^{-1.8}(1.8)^1}{1!}$$

$$= e^{-1.8}(1 + 1.8) = 0.4628$$

$\therefore$ The probability that the number of accidents at most one is 0.4628.

Example 1.73. In a certain factory turning out razor blades, there is a small chance of 0.002 for any blade to be defective. The blades are supplied in packets of 10. Calculate the approximate number of packets containing no defective, one defective and two defective blades in a consignment of 10,000 packets. (Given $e^{-0.02} = 0.9802$).

Solution. Here $N = 10000, p = 0.02, n = 10.$

$\therefore$ Mean $(m) = np = 10 \times 0.002 = 0.02$

Let r be the number of defective blades in a packet.

Let $P(r)$ be the number of packets containing r defective blades, then

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

(*i*) $P(0) =$ Number of packets with no defective blades

$$= 10000 \times \left[0.9802 \times \frac{(0.02)^0}{0!} \right] = 10000\ (0.9802) = 9802$$

$\therefore$ Number of packets with no defective blade = 9802

(*ii*) $P(1) =$ Number of packets with 1 defective blade

$$= 1000 \left[0.9802 \times \frac{(0.02)^1}{1!} \right] = 10000\ [0.9802 \times 0.02]$$

$$= 10000\ (0.019604) = 196.$$

$\therefore$ Number of packets with one defective blade = 196.04 or 196

(*iii*) $P(2)$ = Number of packets with 2 defective blades

$$= 10000\left[0.9802\times\frac{(0.02)^2}{2!}\right]=10000\left[0.9802\times\frac{0.0004}{2}\right]$$

$$= 10000\ (0.00019604) = 1.96$$

$\therefore$ Number of packets with two defective blades = 1.96 or 2.

Example 1.74. If the variance of the Possion distribution is 2, find the probabilities for r, 1, 2, 3, 4 from the recurrence relation of the Possion distribution. Also find $P(x \geq 4)$. (Use $e^{-2} = 0.1353$)

Solution. Here variance = $m = 2$, $P(0) = e^{-2} = 0.1353$ (given)

We known that $$P(r+1) = \frac{m}{r+1}P(r) = \frac{2}{r+1}P(r)$$

Putting $r = 0, 1, 2, 3$ in (*i*) we get

$$P(1) = \frac{2}{0+1}\times P(0) = 2e^{-2} = 2\times 0.1353 = 0.27706$$

$$P(2) = \frac{2}{2}\times P(1) = 0.2706$$

$$P(3) = \frac{2}{3}\times P(2) = \frac{2}{3}\times 0.2706 = 0.1804$$

$$P(4) = \frac{2}{4}\times P(3)\times = \frac{1}{2}\times P(3) = 0.0902$$

Now, $$P(X \geq 4) = 1 - [P(0) + P(1) + P(2) + P(3)]$$

$$= 1 - [0.1353 + 0.2706 + 0.2706 + 0.1804] = 0.1431$$

Example 1.75. If a random variable X follows a Poisson distribution such that $P(X = 2) = 9.\ P(X = 4) + 90\ P(X = 6)$, find the mean and variance of X.

Solution. By Poisson distribution, we have

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

Now $$P(X = 2) = 9.\ P\ (X = 4) + 90\ P(X = 6)$$

$$\Rightarrow \quad \frac{e^{-m}\cdot m^4}{r!} = \frac{9\cdot e^{-m}m^4}{4!} + \frac{90\cdot e^{-m}\cdot m^6}{6!}$$

$$\Rightarrow \quad \frac{3}{4}m^2 + \frac{1}{4}m^4 = 1 \Rightarrow m^4 + 3m^2 - 4 = 0$$

$\Rightarrow (m^2 + 4)\ (m^2 - 1) \Rightarrow m^2 = 1 \Rightarrow m = 1$

Hence, mean = 1 and variance = 1 [$\because$ $m > 0$ and $m^2 + 4 \neq 0$]

[$\because$ Mean = Variance = m]

Example 1.76. An insurance company has discovered that only 0.1% of the population is involved in a certain type of accident every year. If its 1000 policy-holders are selected at random from the population, what is the probability that not more than 5 of its clients are involved in such accident next year? (Use $e^{-1} = 0.3668$)

Solution. Here, n = total number of policy-holders = 1000

Let p = probability of a person being involved in an accident = $\frac{0.1}{100} = 0.001$

$\therefore \quad m = np = 1000 \times 0.001 = 1$

Here, n is large and p is very small.

So, Poisson distribution can be used to find the required probability.

Let X be the Poisson variable with $m = 1$

$\therefore \quad P(X = r) = \frac{e^{-m}m^r}{r!} = \frac{1^r e^{-1}}{r!} = \frac{e^{-1}}{r!}$ $[\because m = 1]$

Required probability = $P(X \leq 5)$

$$= P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)$$

$$= \left(\frac{e^{-1}}{0!} + \frac{e^{-1}}{1!} + \frac{e^{-1}}{2!} + \frac{e^{-1}}{3!} + \frac{e^{-1}}{4!} + \frac{e^{-1}}{5!}\right)$$

$$= e^{-1}\left(1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120}\right)$$

$$= 0.368 \times \frac{326}{120} = \frac{3.68 \times 3.26}{1200} = 0.9997$$

Hence, the required probability is 0.9997.

Example 1.77. Fit a Poisson distribution on the following :

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

(Given that $e^{-0.5} = 0.6065$)

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

m = mean of the distribution

$$= \frac{0\times192 + 1\times100 + 2\times24 + 3\times3 + 4\times1}{192 + 100 + 24 + 3 + 1} = \frac{161}{320} = 0.5 \text{ (approx).}$$

$$P(0) = \frac{e^{-0.5}(0.5)^0}{0!} = 0.6065 \; i.e., f = 320 \times 0.6056 = 194 \text{ (approx)}$$

$$P(1) = \frac{e^{-0.5}(0.5)^0}{1!} = 0.30325 \text{ i.e., } f = 320 \times 0.30325 \times 0.30325 = 97 \text{ (approx)}$$

$$P(2) = \frac{e^{-0.5}(0.5)^1}{2!} = 0.07581 \text{ i.e., } f = 320 \times 0.07581 = 24 \text{ (approx)}$$

$$P(3) = \frac{e^{-0.5}(0.5)^3}{3!} = 0.0126 \text{ i.e., } f = 320 \times 0.0126 = 4 \text{ (approx)}$$

$$P(4) = \frac{e^{-0.5}(0.5)^4}{4!} = 0.0016 \text{ i.e., } f = 320 \times 0.0016 = 0.512 \text{ or } 1 \text{ (approx)}$$

As total number of trials = 320

We have the approximate value as obtained by Poisson distribution as :

x	0	1	2	3	4
f	194	97	24	4	1

Example 1.78. A car hire firm has two cars which it hires out day by day. The number of demands for a car on each day is distributed as poisson distribution with mean 1.5. Calculate the proportion, of days on which (*i*) neither car is used (*ii*) some demand is refused.

Solution. For Poisson distribution

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} \; ; \; x = 0, 1, 2\ 3...$$

Now $\lambda = 1.5$

(*i*) The probability that there is no demand

$$P(X=0) = \frac{e^{-1.5}(1.5)^0}{0!} = e^{-1.5} = 0.223$$

The proportion of days on which neither car is used = 0.2231 = 22.31%

(*ii*) The probability that the demand is refused only when demand is more than 2 cars.

$$P(X>2) = 1 - P(X \le 2)$$
$$= 1 - [P(X=0) + P(X=1) + P(X=2)]$$
$$= 1 - \left\{\frac{e^{1.5}(1.5)^0}{0!} + \frac{e^{-1.5}(1.5)}{0!1} \frac{+e^{1.5}(1.5)^2}{2!}\right\}$$
$$= 0.1912$$

The proportion of days on which some demand is refused = 0.1912 = 19.12%

EXERCISE 1.5

1. If a random variable has a Poisson distribution such that $P(1) = P(2)$, find:

(*i*) Mean of the distribution

(*ii*) $P(4)$

$\left[\textbf{Ans.}(i)2, (ii)\frac{2}{3e^2}\right]$

2. A certain screw making machine produces on average 2 defective screws out of 100, and packs them in boxes of 500. Find the probability of contain 15 defectives screws. [**Ans.** 0.035]

3. Six coins are tossed 1600 times. Use Poisson's distribution to find an expression of getting heads x times. $\left[\textbf{Ans.} P(X=x)=\frac{e^{-25}(25)x}{x!}\right]$

4. In sampling a large number of parts maufactured by machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such sample, how many would be expected to contain at least 3 defective parts? [**Ans.** 324]

5. The probability that a mean aged 50 years will die within a year is 0.01125. What is the probability that of 12 such men at least 11 will reach their fifty first birthday. [**Ans.** 0.9916]

6. Data was collected over a period of 10 years, showing number of deaths from horse kicks in each of the 200 army corps. The distribution of deaths was as following:

Number of deaths	0	1	2	3	4	Total
Frequency	109	65	22	3	1	200

Fit a Poisson distribution to the data the calculate the theoretical frequcies.

$\left[\textbf{Ans.}\ 108.7\frac{(0.61)^r}{k!};109,20,4.1\right]$

7. Find the expectation and variance of a variable X, if the probability $P(x, k)=\frac{e^{-m}m^k}{k!}$ for $k=0, 1, 2, \ldots \infty$ (m is positive real number). [**Ans.** m, m]

8. A manufacture knows that the condensers he makes contain on an average 1% of defective. He packs them in boxes of 100. What is the probability that a box packed at random will contains 3 or more faulty condensers. [**Ans.** 0.08025]

9. Fit a Poisson distribution of the following data and calculate theoretical frequencies:

Deaths	0	1	2	3	4
Frequency	122	60	15	2	1

[**Ans.** Deaths: 0 1 2 3 4
Frequency: 121 61 15 2 1]

10. If the probability that an individual suffers a bad reaction from a certain injection is 0.001. Find the probability that out of 2000 individual: (*i*) exactly 3, (*ii*) more than 2 individuals will suffer a bad reaction. [**Ans.** (*i*) 0.18, (*ii*) 0.32]

11. An insurance company finds that 0.005% of the population dies from a certain kind of accident each year. What is the probability that the company must pay off no more than 3 of 10,000 insured risks against such accident in a given year? [**Ans.** 0.0016]

12. Show that in a Poisson distribution with unit mean, and the mean deviation about the mean is $\frac{2}{e}$ times the standard deviation.

[**Hint:** Mean deviation about the mean $=\Sigma p(r)\,|r-m|$ and $r=0,1,2\ldots.$]

13. If $P(r) = \frac{e^{-m}m^r}{r!}$, prove the relation $P(r+1) = \frac{m}{r+1}P(r)$, $r = 0,1,2,3...\infty$

14. Find the probability that most 5 defective fuses will be found a box of 200 fuses if experience shows that 2% of such fuses are defective. [**Ans.** 0.7845]

15. Fill in the blanks :
    (*i*) If a random varible x follows Poisson distribution such that $P(x=1) = P(x=2)$, the mean of the distribution is ...
    (*ii*) If the standard deviation of Poisson distribution is $\sqrt{2}$, the probability for $r = 2$ is ....
    (*iii*) If the probability of defective fuse is 0.05, the variance for the distribution of defective fuses in a total of 40 is .....
    (*iv*) If X is a Poisson variate such that $P(2) = 9P(4) + 90\,P(6)$, the value of σ is .....

    [**Ans.** (*i*) 2, (*ii*) $2\left(\frac{1}{e}\right)^2$, (*iii*) 2, (*iv*) 1]

16. In a certain factory 2% of the items produced are defective, the items are packed in boxes of 100 items. What is the probability that there will be: (*i*) 2 defective items, (*ii*) at least three defective items and (*iii*) 2 < defective items < 5. [**Ans.** (*i*) 0.272, (*ii*) 0.32, (*iii*) 0.272]

17. Suppose 2% of the people on an average are left-handed. Find, (*i*) the probability of the finding 3 or more left-hunded, (*ii*) probability of finding none of one left-handed. [**Ans.** (*i*)0, (*ii*) 0.998]

18. The number of monthly breakdowns of a computer is a random variable having Poisson distribution with mean equal to 1.8. Find the probability of that the computer will function for a month: (*i*) without a breakdown, (*ii*) with only one breakdown and (*iii*) with at least one.

    [**Ans.** (*i*) 0.16529, (*ii*) 0.2975, (*iii*) 0.8347]

1.31. MOMENT OF RANDOM VARIABLE

The "moments" of a random variable (or its distribution) are expected values of powers or related functions of the random variable.

The r^{th} moment of X is $\mu'_r = E(X^r) = \Sigma x^r P(X = x)$

In particular, the first moment is mean, $\mu_X = E(X) = \mu'_1$

The r^{th} central moment of X is $\mu_r = E(X - \mu_X)^r$

In particular, the second central moment is the variance, $\sigma_X^2 = \text{Var}\,X = E(X - \mu_X)^2$

1.32. RELATION BETWEEN r^{th} MOMENT OF RANDOM VARIABLE AND r^{th} CENTRAL MOMENT OF RANDOM VARIABLE

$$\mu_1 = E(X - \mu_X)^1 = E(X) - E(\mu_X)$$
$$= \mu_X - \mu_X = 0$$
$$\mu_2 = E(X - \mu_X)^2 = E(X^2 + \mu_X^2 - 2X\mu_X)$$
$$= E(X^2) + \mu_X^2 - 2\mu_X E(X)$$
$$= \mu_2^1 + \mu_X^2 - 2\mu_X^2$$
$$= \mu_2^1 - \mu_1'^2$$

[$\because$ first central moment is mean]

$$\mu_3 = E(X-\mu_X)^3 = E(X^3 + \mu_X^3 - 3\times {}^2\mu_X + 3\mu_X{}^2X)$$
$$= E(X^3) - \mu_X^3 - 3\mu_X E(X^2) + 3\mu_X^2 E(X)$$
$$= E(X^3) - \mu_1'^3 - 3\mu_1'\mu_2' + 3\mu_1'^2\mu_1' \qquad [\because \mu_X = \mu_1]$$
$$= \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3$$
$$\mu_4 = E(X-\mu_X)^4 = E(X^4 + \mu_X^4 + 6X^2\mu_X^2 - 4X^3\mu_X - 4X\mu_X^3)$$
$$= E(X^4) + \mu_X^4 + 6\mu_X^2 E(X^2) - 4\mu_X E(X^3) - 4\mu_X^3 E(X)$$
$$= \mu_4^1 + (\mu_1^1)^4 + 6\mu_1'^2\mu_2^- - 4\mu_1^1\mu_1^3 - 4'^4_1$$
$$= \mu_4^1 + 6\mu_1'^2\mu_2' - 4\mu_1'\mu_2' - 3\mu_1'^4$$

SOLVED EXAMPLES

Example 1.79. Let X be a discrete random variable having probability mass function.

$$P_X(x) = \begin{cases} 1/2 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 1/6 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

Find the third moment of X.

Solution. Thid moment $= E(X^3)$

$$= \Sigma x^3 P_X(x)$$
$$= \frac{1}{2}\cdot 1^3 + \frac{1}{3}\cdot 2^3 + \frac{1}{6}\cdot 3^3$$
$$= \frac{1}{2} + \frac{8}{3} + \frac{27}{6} = \frac{23}{3}$$

Example 1.80. Let X be a discrete random variable with probability mass function.

$$P_X(x) = \begin{cases} 3/4 & \text{if } x = 1 \\ 1/4 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Find the third central moment of X.

Solution. $E(X) = \mu_X = \Sigma x P_X(x)$

$$= 1\times\frac{3}{4} + 2\times\frac{1}{4} = \frac{5}{4}$$

The third central moment of X can be computed as follows:

$$E(X-\mu_X)^3 = \sum\left(x - \frac{5}{4}\right)^3 P_X(x)$$

$$= \left(1 - \frac{5}{4}\right)^3 \times\frac{3}{4} + \left(2 - \frac{5}{4}\right)^3 \times\frac{1}{4}$$

$$= \left(\frac{-1}{4}\right)^3 \cdot \frac{3}{4} + \left(\frac{3}{4}\right)^3 \cdot \frac{1}{4}$$

$$= \frac{3}{32}$$

Example 1.81. Find the first four moments of Binomial distribution about mean.

Solution. The Binomial distribution is given by

$$P(X = x) = {}^nC_x p^x q^{n-x}, x = 0,1,2 \ldots, n \qquad \text{where } p + q = 1$$

$$\therefore \quad \mu_1^1 = \sum_{x=0}^{n} xP(X = x) = \sum_{x=0}^{n} x\, {}^nC_x p^x q^{n-x}$$

$$= {}^nC_1 pq^{n-1} + 2{}^nC_2 p^2 q^{n-2} + \ldots + x^n C_n p^n$$

$$= np(q^{n-1} + (n-1)pq^{n-2} \ldots + p^{n-1})$$

$$= np(q + p)^{n-1} = np$$

$$\mu_1' = np$$

Now $$\mu_2' = \sum_{x=0}^{n} x^2 P(X = x)$$

$$= \sum_{x=0}^{n} [x(x-1) + x]\, {}^nC_x p^x q^{n-x} \qquad \ldots(i)$$

$$= \sum_{x=0}^{n} x(x-1)^n C_x p^x q^{n-x} + \sum_{x=0}^{n} x \cdot {}^nC_x p^x q^{n-x}$$

$$= [2.1\, {}^nC_2\, p^2 q^{n-2} + 3.2\, {}^nC_3 p^3 q^{n-3} \ldots + p^{n-2}] + E(X)$$

$$= n(n-1)p^2\, [q^{n-2} + (n-2)pq^{n-3} \ldots + p^{n-2}] + np$$

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

$$\mu_2' = n(n-1)p^2 + np \qquad \ldots(ii)$$

$$\therefore \quad \mu_2 = \mu_2' - (\mu_1')^2 = n(n-1)p^2 + np - n^2p^2 \qquad [\because \quad p + q = 1]$$

$$= np(1-p) = npq$$

$$\mu_1' = \sum_{x=0}^{n} x^3\, P(X = x)$$

$$\mu_2' = \sum_{x=0}^{n} \{x + 3x(x-1) + x(x-1)\,(x-2)\}\, {}^nC_x p^x q^{n-x} \qquad \ldots(iii)$$

Then $$\mu_3' = np + 3n(n-1)p^2 + n(n-1)\,(n-2)p^3 \quad \text{[Using } (i) \text{ and } (ii)\text{]} \qquad \ldots(iv)$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$= np + 3n(n-1)p^2 + n(n-1)\,(n-2)p^3 - 3\{n(n-1)p^2 + np\}\, np + 2n^3p^3$$

$$= npq((1-p) - p) = npq(q - p)$$

Now $$\mu_4' = \sum_{x=0}^{n} x^4 \, P(X = x)$$

$$= \sum_{x=0}^{n} \{x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x\} \, P(X=x)$$

From (*iii*) and (*iv*) we obtain

$$\mu_4' = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= \{n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np\} - 4\{n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np\}np + 6\{n(n-1)p^2 + np\}\, n^2p^2 - 3n^4p^4$$

Solving

$$\mu_4 = npq\,\{1 + 3pq(n-2)\}.$$

Example 1.82. Deduce the first four moments about mean of the Poisson distribution from those of Binomial distribution.

Solution. Poisson distribution is the limiting form of binomial distribution when $n \to \infty, p \to 0$ and $np = m$ (a finite constant)

For binomial constant

$$\mu_1' = np, \quad \mu_2' = npq, \quad \mu_3' = npq\,(p-q) \quad \text{and} \quad \mu_4' = npq\{1 + 3pq\,(n-2)\}$$

For Poisson distribution

$$\mu_1' \text{ for P.D.} = \lim_{n\to\infty} (np) = m$$

$$\mu_2 \text{ for P.D.} = \lim_{n\to\infty} (npq) = \lim_{n\to\infty} np(1-p)$$

$$= \lim_{n\to\infty} m\left(1 - \frac{m}{n}\right) = m \qquad [\because \quad np = m]$$

$$\mu_3 \text{ for P.D.} = \lim_{n\to\infty} npq(q-p)$$

$$= \lim_{n\to\infty} npq(1-2p) = \lim_{n\to\infty} m\left(1 - \frac{m}{n}\right)\left(1 - \frac{2m}{n}\right)$$

$$= m$$

$$\mu_4 \text{ for P.D.} = \lim_{n\to\infty} npq\{1 + 3pq\,(n-2)\}$$

$$= \lim_{n\to\infty} m\left\{1 - \frac{m}{n}\right\}\left\{1 + \frac{3m}{n}\left(1 - \frac{m}{n}\right)(n-2)\right\}$$

$$= \lim_{n\to\infty} m\left\{1 - \frac{m}{n}\right\}\left\{1 + 3m\left(1 - \frac{m}{n}\right)\left(1 - \frac{2}{n}\right)\right\} = 3m^2 + m.$$

1.33. MOMENT GENERATING FUNCTION

The moment generating function (m.g.f.) of a random variable X having the probability function $f(x)$ is given by

$$M_x(t) = E(e^{tX}) \qquad \ldots(i)$$

$$= \int e^{tx} f(x)\, dx \qquad \text{(for continuous probability distribution)}$$

$$= \sum_x e^{tx} f(x) \qquad \text{(for discrete probability distribution)}$$

where t is a real constant. Thus

$$M_x(t) = E(e^{tX}) = E\left[1 + tX + \frac{t^2X^2}{2!} + \ldots + \frac{t^rX^r}{r!} + \ldots\right]$$

$$= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \ldots\ldots + \frac{t^r}{r!}E(X^r) + \ldots\ldots$$

$$= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots\ldots + \frac{t^r}{r!}\mu_r' + \ldots\ldots \qquad \ldots(ii)$$

The coefficient of $\frac{t^r}{r!}$ in the expansion (ii) is μ_r', the rth moment of X about origin. Since $M_X(t)$ generates moments, it is known as moment generating function.

Note: Differentiating (ii) w.r.t. t and putting $t = 0$, we get

$$\frac{d}{dt}[M_X(t)]_{t=0} = \left(\mu_1' + t\mu_2' + \frac{t^2}{2!}u_3' + \ldots\ldots\right)_{t=0} = \mu_1'$$

Similarly, $$\frac{d^2}{dt^2}[M_X(t)]_{t=0} = \mu_2'$$

and $$\frac{d^r}{dt^r}[M_X(t)] = \mu_r' \qquad \ldots(iii)$$

In general, the moment generating function of X about the point $x = a$ is defined as

$M_X(t)$ (about $x = a$) $= E[e^{t(x-a)}]$

$$= E\left[1 + t(X-a) + \frac{t^2}{2!}(X-a)^2 + \ldots\ldots + \frac{t^r}{r!}(X-a)^r + \ldots\ldots\right]$$

$$= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots\ldots + \frac{t^r}{r!}\mu_r' + \ldots\ldots \qquad \ldots(iv)$$

where $\mu_r' = E[(X-a)^r]$, is the rth moment about the point $x = a$

Similarly, $M_X(t)$ (about mean) $= E(e^{t(X-\bar{x})})$

$$= E\left[1 + t(X-\bar{x}) + \frac{t^2}{2}(X-\bar{x})^2 + \ldots + \frac{t^r}{r!}(X-\bar{x})^r + \ldots\right]$$

$$= 1 + t\mu_1 + \frac{t^2}{2!}\mu_2 + \ldots + \frac{t^r}{r!}\mu_r + \ldots$$

where $\mu_r = E[(x-\bar{x})^r]$, is the rth moment about mean $(\bar{X})$.

1.34. PROPERTIES OF MOMENT GENERATING FUNCTION

1. $M_{cX}(t) = M_X(ct)$, c being a constant.

Proof: By definition

LHS $= M_{cX}(t) = E(e^{tcx})$

$$= E(e^{(ct)x})$$
$$= M_X(ct).$$

2. The moment generating function of the sum of a number of independent random variables is equal to the product of their respective moment generating function. *i.e.*,

If $x_1, x_2, \ldots\ldots, x_n$ are independent random variables then the moment generating function of their sum $x_1 + x_2 + \ldots + x_n$ is given by

$$M_{X_1 + X_2 + \ldots + X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \ldots\ldots M_{X_n}(t)$$

Proof: By definition,

$$M_{X_1 + X_2 + \ldots + X_n}(t) = E\left[e^{t(X_1 + X_2 + \ldots + X_n)}\right]$$
$$= E\left[e^{tX_1 + tX_2 + \ldots + tX_n}\right]$$
$$= E\left[e^{tX_1} \cdot e^{tX_2} \ldots e^{tX_n}\right]$$
$$= E(e^{tX_1}) \cdot E(e^{tX_2}) \ldots\ldots E(e^{tX_n})$$

$[\because X_1, X_2, \ldots, X_n$ are independent$]$

$$= M_{X_1}(t) \cdot M_{X_2}(t) \ldots\ldots M_{X_n}(t)$$

Hence, the theorem.

1.35. EFFECT OF CHANGE OF ORIGIN AND SCALE ON M.G.F.

Let X be a random variable with $M_X(t) = E(e^{tX})$

Let $U = \dfrac{X-a}{h}$, where a and h are constants

$\Rightarrow$ $$X = a + hU$$

$$M_U(t) = E(e^{tU}) = E(e^{t(X-a)/h})$$

Then,

$$M_U(t) = e^{-at/h} E\left(e^{\frac{X}{h}}\right)$$

$$\Rightarrow \quad M_U(t) = e^{-at/h} M_X\left(\frac{t}{h}\right) \qquad ...(i)$$

$$\Rightarrow$$

In other form

$$M_X(t) = E(e^{tX})$$
$$= E[e^{t(a + hU)}]$$
$$= e^{at} E(e^{thu})$$
$$M_X(t) = e^{at} M_U(th) \qquad ...(ii)$$

SOLVED EXAMPLES

Example 1.83. Find the m.g.f. of a random variable whose moments are

$$\mu_r' = (r+1)! \, . \, 2^r.$$

Solution. The m.g.f. is given by

$$M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_r' = \sum_{r=0}^{\infty} \frac{t^r}{r!} (r+1)! \, 2^r$$

$$= \sum_{r=0}^{\infty} (r+1) (2t)^r$$

$$M_X(t) = 1 + 2.(2t) + 3(2t)^2 + 4(2t)^3 + ...$$
$$= (1 - 2t)^{-2}.$$

Example 1.84. Let the random variable X assume the value 'r' with the probability function given by

$$P(x = r) = q^{r-1} p, \quad r = 1, 2, 3, ...$$

Find the m.g.f. of X and hence its mean and variance.

Solution. $M_X(t) = E(e^{tX})$

$$= \sum_{r=1}^{\infty} (e^{tr} \, q^{r-1} p) = \frac{p}{q} \sum_{r=1}^{\infty} (qe^t)^r$$

$$= \frac{p}{q} . qe^t \sum_{r=1}^{\infty} (qe^t)^{r-1}$$

$$= pe^t [1 + qe^t + (qe^t)^2 + ...]$$

$$= \frac{pe^t}{1 - qe^t}$$

$\therefore$

$$\mu_1' = \frac{d}{dt} [M_X(t)] = \frac{d}{dt} \left[\frac{pe^t}{1 - qe^t}\right]_{t=0}$$

$$= \left[\frac{(1 - e^t q)(pe^t) + (pe^t)(qe^t)}{(1 - qe^t)^2} \right]_{t=0}$$

$$= \frac{pe^t}{(1 - qe^t)^2} \Bigg]_{t=0} = \frac{1}{p}$$

and
$$\mu_2' = \frac{d^2}{dt^2} [M_X(t)]_{t=0}$$

$$= \frac{d}{dt} \left[\frac{pe^t}{(1 - qe^t)^2} \right]_{t=0} = \frac{pe^t(1 + qe^t)}{(1 - qe^t)^3} \Bigg]_{t=0}$$

$$= \frac{1+q}{p^2}$$

Hence,
$$\text{Mean} = \mu_1' \text{ (about origin)} = \frac{1}{p}$$

and
$$\text{variance} = \mu_2 = \mu_2' - \mu_1'^2$$

$$= \frac{1+q}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}.$$

Example 1.85. Show that the m.g.f. of a random variable X having the probability density function

$$f(x) = \begin{cases} \frac{1}{3} & , \quad -1 < x < 2 \\ 0 & , \quad \text{elsewhere} \end{cases}$$

is
$$M_X(t) = \begin{cases} \frac{e^{2t} - e^{-t}}{3t} & , \quad t \neq 0 \\ 1 & , \quad t = 0 \end{cases}$$

Solution.
$$M_X(t) = E(e^{tX})$$

$$= \int_{-1}^{2} e^{tx} f(x)\, dx = \int_{-1}^{2} \frac{1}{3} e^{tx}\, dx$$

$$= \frac{1}{3t}[e^{tx}]_{-1}^{2} = \frac{1}{3t}(e^{2t} - e^{-t}), \quad t \neq 0$$

However, when $t = 0$, then

$$M_X(t) = \int_{-1}^{2} \frac{1}{3} e^0\, dx = \int_{-1}^{2} \frac{1}{3}\, dx = \left(\frac{x}{3} \right)_{-1}^{2}$$

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

Thus,
$$M_x(t) = \frac{e^{2t} - e^t}{3t}, \quad t \neq 0$$
$$= 1, \; t = 0.$$

Example 1.86. Find the m.g.f. of the random variable X having the probability density function

$$f(x) = \begin{cases} x & 0 \le x < 1 \\ 2 - x & 1 \le x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Also, find the mean and variance of X using m.g.f.

Solution.
$$M_X(t) = E(e^{tX})$$
$$= \int_0^1 e^{tx}\, x\, dx + \int_1^2 e^{tx}(2 - x)\, dx$$
$$= \left(\frac{e^{tx}}{t} \cdot x\right)_0^1 - \int_0^1 \frac{e^{tx}}{t} \cdot 1\, dx + \left[\frac{e^{tx}}{t}(2 - x)\right]_1^2 - \int_1^2 \frac{e^{tx}}{t}(-1)\, dx$$
$$= \frac{e^t}{t} - \left(\frac{e^{tx}}{t^2}\right)_0^1 + \frac{-e^t}{t} + \left[\frac{e^{tx}}{t^2}\right]_1^2$$
$$= \frac{e^t}{t} - \left(\frac{e^t}{t^2} - \frac{1}{t^2}\right) - \frac{e^t}{t} + \left[\frac{e^{2t}}{t^2} - \frac{e^t}{t^2}\right] = \frac{e^{2t}}{t^2} - \frac{2e^t}{t^2} + \frac{1}{t^2}$$
$$= \frac{1}{t^2}(e^{2t} - 2e^t + 1) \qquad \ldots(i)$$
$$= \frac{(e^t - 1)^2}{t^2} \qquad \ldots(ii)$$

Now expanding $M_X(t)$ given in (i), we have

$$M_x(t) = \frac{1}{t^2}\left[\left(1 + 2t + \frac{(2t)^2}{2!} + \frac{(2t)^3}{3!} + \frac{(2t)^4}{4!} + \ldots\right) - 2\left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \ldots\right) + 1\right]$$
$$= \frac{1}{t^2}\left[t^2 + t^3 + \frac{7}{12}t^4 \ldots\right]$$

$\Rightarrow$
$$M_X(t) = 1 + t + \frac{7}{12}t^2 + \ldots\ldots \qquad \ldots(iii)$$

Hence,
$$\mu_1' = \text{Coefficient } t \text{ in } M_X(t) = 1 = \text{Mean}$$
$$\mu_2' = \text{Coefficient } \frac{t^2}{2!} \text{ in } M_X(t) = 2!\, \frac{7}{12} = \frac{7}{6}$$

$\therefore$
$$\text{Variance } (\mu_2) = \mu_2' - \mu_1'^2$$
$$= \frac{7}{6} - (1)^2 = \frac{1}{6}.$$

EXERCISE 1.6

1. Define moment generating function? Why is it called moment generating function?
2. If $M(t)$ is the m.g.f. of a random variable X about the origin, show that the moment μ_r' is given by

$$\mu_r' = \left[\frac{d^r M(t)}{dt^r}\right]_{t=0}.$$

3. A random variable X has probability function $p(x) = \frac{1}{2^x}$; $x = 1, 2, 3, \ldots$.

Find its m.g.f., mean and variance. $\left[\text{Ans. } \frac{e^t}{2-e^t}, 2, 2\right]$

4. Find the moment generating function of the random variable X whose probability density function is

$$f(x) = \begin{cases} \frac{x}{2}, & 0 \le x \le 2 \\ 0, & \text{otherwise} \end{cases} \qquad \left[\text{Ans. } \frac{e^{2t}}{t}\left(1 - \frac{1}{2t}\right)\right]$$

5. Find the m.g.f. of a random variable X whose probability function is given by

$$P(X = x) = p(1-p)^x, \quad x = 0, 1, 2, \ldots. \infty \qquad [\text{Ans. } M_X(t) = p(1 - qet)^{-1}]$$

6. The probability density function of the random variable X follows the following probability law:

$$p(x) = \frac{1}{2\theta} \exp\left(-\frac{|x-\theta|}{\theta}\right), \quad -\infty < x < \infty$$

Find m.g.f. of X. Hence or otherwise find $E(X)$ and $V(X)$. [Ans. $e^{\theta t}(1 - \theta^2 t^2)^{-1}$; $\theta, 2\theta^2$]

7. Find the m.g.f. for $f(x) = ce^{-cx}$, $c > 0$, $0 \le x < \infty$ and deduce that $\mu_2 = \frac{1}{c^2}$. $\left[\text{Ans. } \left(1 - \frac{t}{c}\right)^{-1}\right]$

8. If the m.g.f. of a random variable X is of the form $(0.4 + 0.6\, e^t)^3$, show that the m.g.f. of a random variable $Y = 3X + 2$ is $(0.4 + 0.6\, e^{3t})^3$.

1.36. MULTINOMIAL DISTRIBUTION

The distribution can be regarded as a generalisation of Binomial distribution.

When there are more than two mutually exclusive outcomes of a trial, the observations lead to multinomial distribution. Suppose $E_1, E_2, \ldots\ldots, E_k$ are k mutually exclusive and exhaustive outcomes of a trial with the respective probabilities $p_1, p_2, \ldots\ldots, p_k$.

The probability that E_1 occurs x_1 times, E_2 occurs x_2 times... and E_k, occurs x_k times in n independent observations, is given by

$$p(x_1, x_2, \ldots, x_k) = c p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

where $\Sigma x_i = n$ and c is the number of permutation of the events $E_1, E_2, \ldots, E_k$.

between the two variables. Furthermore, there should not be any blank in one series corresponding to the given value in the other series.

In a bivariate distribution, we are given a set of pair of observations. We are interested to find relationship, if any, between the two variables. *Correlation is a statistical tool which studies the relationship between two variable, correlation analysis involves various methods and techniques used for studying and measuring the extend of the relationship between two variables.*

4.21. TYPES OF CORRELATION

Correlation is classified in the following three ways:

1. Positive and negative correlation
2. Simple, partial or multiple correlation
3. Linear and non-linear correlation

4.21.1. Positive and Negative Correlation

Positive correction: If the value of the two variables deviate in the same direction *i.e.*, an increase in the value of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the value of one variable say X results on an average a corresponding decrease in the other variable say Y, the correlation is said to be **positive** or **direct**. The height and weight of a growing child is an example of positive correlation.

Negative correlation: The correlation is said to be **negative** or **inverse** if the two variables X and Y deviate in the opposite direction, *i.e.*, if the increase (or decrease) in the values of the variable X results, on the average, in a corresponding decrease (or increase) in the values of other variable Y. The price and demand of a commodity or the consumption of electricity and its bill etc., are examples of negative or inverse correlation.

4.21.2. Simple, Partial and Multiple Correlation

When only two variables are studied, it would be a case of simple correlation. When more than two variables are involved then the problem, may be of either partial or multiple correlation.

In ***partial correlation***, we measure the correlation between two variables (one dependent and other independent variable) when all other variable involved are kept constant. For example, if we limit our correlation analysis of yield and rainfall to periods where a certain average temperature existed, it will be a case of partial correlation.

In ***multiple correlation,*** three or more than three variable are studied simultaneously. For example, if we study yield of wheat per acre and both the amounts of rainfall and fertilizers used, it is a problem of multiple correlation.

4.21.3. Linear and Non-linear Correlation

The distinction between linear and non-linear correlation is based upon the consistancy of the ratio of change between the variables. If the amount of change in one variable tends to be a constant ratio to the amount of change in the other variable than the correlation is said to be linear.

Correlation would be called non-linear or curvilinear if the amount of change in one variable does not bear a constant ratio to the amount of changes in the other variables.

The following two diagrams will illustrate the differences between linear and curlilinear correlation:

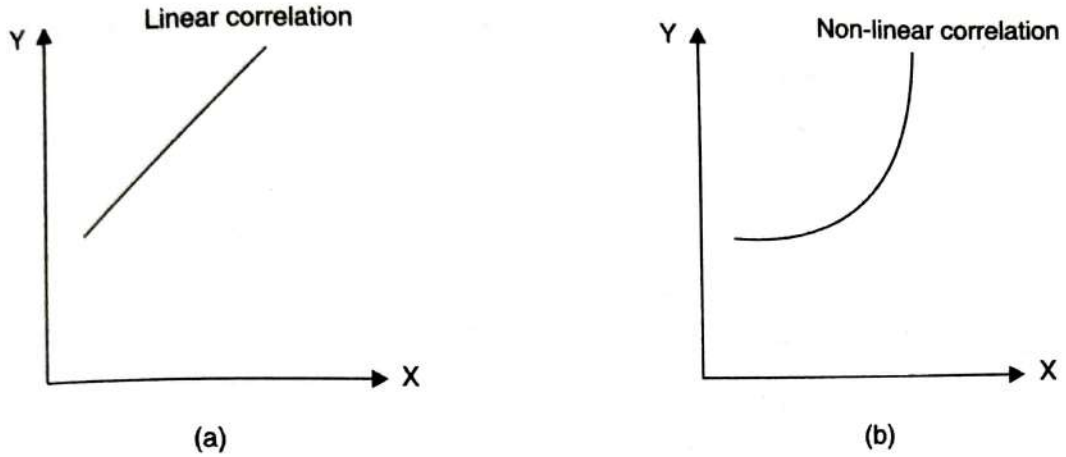

Fig. 4.1

4.21.4. Significance or the Study of Correlation

The study of correlation is of great significance because of the following reasons:

1. Most of the variables show some kind of relationship. With the help of correlation analysis is the degree of correlation among them can be measured in one figure only.
2. Sometimes the existence of relationship between two or more variables enable us to predict what will happen is future. For example, if production of sugar increases, other things remaining the same, we may expect a fall in its price.
3. Once we know that two variables are correlated, we can estimate the value of one variable given the value of another. This is done with the help of regression equations.
4. In business, it enables the businessman to estimate costs, sales and prices etc. It facilitates decision-making in business. It may suggest expected changes in one variable on the basis of change in the other. It reduce the risk of uncertainty in decision-making,

However, if the concept of correlation is employed carelessly, its analysis may lead to misleading conclusions. Hence, one has to be very cautions while making use of it.

4.21.5. Coefficient of Correlation

Coefficient of correlation is calculated to study the extent or the degree of correlation that exists between two variables. Correlation coefficient of the variables x and y will be denoted by r_{xy} or $r(x, y)$ or $\rho(x, y)$ or ρ.

4.21.6. Properties of Coefficient of Correlation

(*i*) It is a measure of the closeness of a fit in a relative sense.

(*ii*) Correlation coefficient lies between -1 and $+1$, *i.e.*, $-1 \leq r \leq 1$.

(*iii*) The correlation is perfect and positive if $r = 1$ and it is perfect and negative if $r = -1$.

(*iv*) If $r = 0$, then there is not correlation between the two variables and thus the variables are said to be independent.

(*v*) The correlation coefficient is a pure number and is not affected by a change of origin and scale in magnitude.

This property state that if the original pair of variables x and y is changed to a new pair of variables u and v by affecting a change or origin and scale for both x and y *i.e.*,

$$u = \frac{x-a}{b} \text{ or } u = \frac{x}{b} - \frac{a}{b};$$

and

$$v = \frac{y-c}{d} \text{ or } v = \frac{y}{d} - \frac{c}{d},$$

where a and c are the origins of x and y and b and d are respective scales and then we have:

$$r_{xy} = \frac{bd}{|b||d|} r_{uv} \qquad ...(i)$$

r_{xy} and r_{uv} being the coefficient of correlation between 'x and y' and 'u and v' respectively. The equation (*i*) shows that numerically, the two correlation coefficient remain equal and they would have opposite signs only when b and d, the two scales, differ in sign.

(*vi*) It is a relative measure of association between two or more variables.

4.21.7. Covariance

Let the n pairs observations on two quantitative variables X and Y be

$$(x_1, y_1)\ (x_2, y_2), \ldots\ldots, (x_n, y_n)$$

where $x_1, x_2, \ldots$ denote observed values of the variable X and $y_1, y_2 \ldots$ those of Y.

The deviation of the observed X values of their mean $\bar{x}$, are

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots\ldots, x_n - \bar{x}$$

and the deviation of the observed Y values from their mean y, are

$$y_1 - \bar{y}, y_2 - \bar{y}, \ldots\ldots, y_n - \bar{y}$$

The covariance of X and Y is the sum of n products

$$(x_1 - \bar{x})\ (y_1 - \bar{y}), (x_2 - \bar{x})\ (y_2 - \bar{y}), \ldots\ldots, (x_n - \bar{x})\ (y_n - \bar{y})$$

divided by n.

$$\therefore \quad \text{Cov}\,(X, Y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

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

Thus

$$\boxed{\text{Cov}\,(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$$

Now $$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i\, y_i - n\overline{xy}$$

$$= \sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

Thus to find the covariance of n pairs of observations (x_1, y_2), (x_2, y_2), ..., (x_n, y_n), proceed as follows:

(*i*) Obtain the sums $\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i$

(*ii*) Obtain the sum $\sum_{i=1}^{n} x_i\, y_i$ of the products $x_i\, y_i$.

(*iii*) Calculate the difference $\sum x_i\, y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)$ and divide by n.

SOLVED EXAMPLES

Example 4.74. Calculate the covariance of the following pairs of observations of the two variates X and Y.

(1, 6), (2, 9), (3, 6), (4, 7), (5, 8), (6, 5), (7, 12), (8, 3), (9, 17), (10, 1).

Solution. Here $$\sum x_i = 1+2+3+4+5+6+7+8+9+10 = 55.$$

$$\sum y_i = 6+97+6+7+8+5+12+3+17+1 = 74.$$

$$\sum x_i\, y_i = 1\times 6 + 2\times 9 + 3\times 6 + 4\times 7 + 5\times 8 + 6\times 5 + 7\times 12 + 8\times 3 + 0\times 17 + 10\times 1$$

$$= 6 + 18 + 18 + 28 + 40 + 30 + 84 + 24 + 153 + 10 = 411.$$

$\therefore$ $$\text{Cov.}(X, Y) = \frac{1}{n}\left(\sum x_i\, y_i - \frac{1}{n}\sum x_i \sum y_i\right)$$

$$= \frac{1}{10}\left(411 - \frac{1}{10}\times 55\times 74\right) = \frac{1}{10}(411 - 407) = 0.4$$

Example 4.75. Calculate the covariance of the following pairs of observations of the variables X and Y.

(15, 44), (20, 43), (25, 45), (30, 37), (40, 34), (50, 37).

Solution. Here $$\bar{x} = \frac{15+20+25+30+40+50}{6} = \frac{180}{6} = 30;$$

$$\bar{y} = \frac{44+43+45+37+34+37}{6} = \frac{240}{6} = 40$$

Here $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x - 30)(y - 40) = 60 - 30 - 25 + 0 - 60 - 60 = -235; n = 6.$

$\therefore$ $\text{Cov}(X, Y) = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \dfrac{-235}{6} = -39.17$ nearly.

4.21.8. Methods of Studying Correlation

The various methods of ascertaining whether two variables are correlated or not are:

(*i*) Scatter diagram method

(*ii*) Karl Pearson's coefficient of correlation method

(*iii*) Spearman's Rank correlation coefficient method

4.21.8.1. Scatter Diagram

Scatter diagram is a special type of dot chart. It is a useful technique for examining the form of relationship visually without calculating any numerical value. In this method, the given data

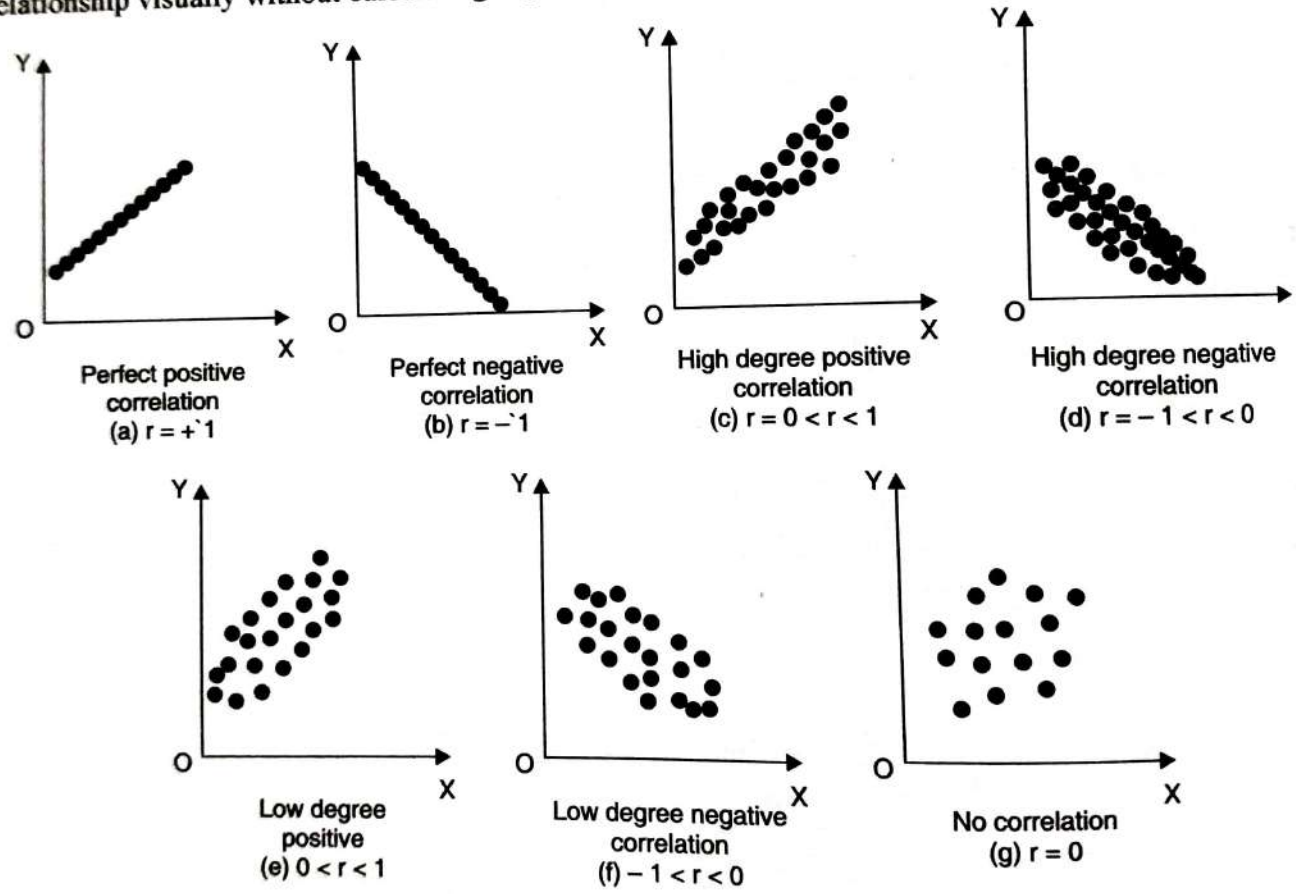

Fig. 4.2

are plotted on a graph-paper in form of dots. For each pair of X and Y values, we put a dot (or a point) and, in this way we obtain any dots equal in number of observations. If a statistical series shows 10 pairs of X and Y values we will have 10 dots. The whole set of dots is called Scatter Diagram. The term scatter refers to the dispersion spread of dots on the graph. Scatter diagram is also known as dot diagram. By mere observation, we can have an idea so as to whether correlation is positive or negative. If the trend of the dotted points is upward to the right, then correlation is positive. On the other hand, if the trend of

dots move from upper-hand corner to lower right-hand side, then the correlation is negative. The greater the scatter of points on the graph, the lesser is the correlation between the two variables. On the other hand, the more closely points come to falling on a line, the higher the degree of relationship. It may be noted that scatter diagram does not show the exact measurement of the relationship. It only gives us an idea of the relationship between two sets of data. Let us now take some examples of scatter diagrams.

(*i*) If all the points lie on a straight line from the lower left-hand corner to the upper right-hand corner, correlation is said to be perfectly positive (*i.e.*, $r = +1$). *See* Fig. 4.2(*a*).

(*ii*) If all the points are lying on a straight line rising from the upper-left hand corner to the lower-right hand corner of the diagram correlation will said to be perfectly negative (*i.e.*, $r = -1$). See Fig. 4.2(*b*).

(*iii*) If the plotted points are very close to each other it shows high degree correlation Fig. 4.2(*c*) shows high degree positive correlation whereas Fig. 4.2(*d*) indicates high degree negative correlation.

(*iv*) If the plotted points are not very close to each but show some downward or upward bend, there is low degree correlation Fig. 4.2(*e*) shows low degree positive correlation and Fig. 4.2(*f*), depicts low degree negative correlation.

(*v*) If the plotted points show no trend at all, it shows absence of correlation between the two variables [See Fig. 4.2(*g*)].

Merits and Demerits of Scatter Diagram

Mertis:

(*i*) It is a simple and non-mathematical method of studying correlation.

(*ii*) It is not influence by the size of extreme values whereas most of the mathematical methods lack this quality.

(*iii*) It is the first step in investigating the relationship between two variables.

(*iv*) In case of linear relationship, it gives a clear picture of the proportionate change in the value of one variable for a change in the other.

Demerits:

(*i*) The exact degree of correlation (in numerical terms) cannot the obtained from it. It only gives a rough estimate about the relationship between the two variables.

(*ii*) It only shows the type of correlation between the two variables.

(*iii*) It is not possible to draw a scatter diagram on a graph paper in case of more than two variables.

4.21.8.2. Karl Pearson's Coefficient of Correlation

Coefficient of correlation is calculated to study the extent or the degree of correlation that exists between two variables.

Definition: The correlation coefficient $\rho(X, Y)$ between two variables X and Y is given by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X), \text{Var}(Y)}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right\}}}$$

$$\rho(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}\left\{\sum_{i=1}^{n}(y_i - \bar{y})^2\right\}}}$$

Denominator in the definition of ρ will always be positive, so that the value of ρ is positive or negative according as the covariance value is positive or negative. Its value lies between -1 and $+1$.

i.e., $$-1 \le \rho(X, Y) \le 1.$$

This is known as Karl Pearson's coefficient of correlation as it was first introduced by Karl Pearson.

We know that $\sum(x_i - \bar{x})^2$ and $\sum(y_1 - \bar{y})^2$ are more easily calculated as

$$\sum x_1^2 - \frac{\left(\sum x_i\right)^2}{n} \text{ and } \sum y_1^2 - \frac{\left(\sum y_i\right)^2}{n}$$

Also $\sum(x_i - \bar{x})(y_i - \bar{y})$ is more easily calculated as

$$\sum x_i\, y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

Thus,
$$\rho(X, Y) = \frac{\sum x_i\, y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\left\{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right\}\left\{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right\}}}$$

or
$$\rho(X, Y) = \frac{n\sum x_i\, y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

Also
$$\rho(X, Y) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

If
$$u_i = \frac{x_i - a}{h},\ v_i = \frac{y_i - b}{k},\ \text{then}$$

$$\rho(X, Y) = \rho(U, V) = \frac{\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{\sqrt{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2}\sqrt{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2}}$$

The value of ρ is independent of the origin and the scale.

Note. If $\bar{x}$ is a whole number, take $a = \bar{x}$.

If $\bar{y}$ is whole number, take $b = \bar{y}$.

[Sum of deviations from mean is always zero].

Remember:

1. (*i*) If $\rho(X, Y) = 1$, we say that there is a perfect positive correlation between x and y.
   (*ii*) If $\rho(X, Y) = -1$, we say that there is a perfect negative correlation between x and y.
   (*iii*) If $\rho(X, Y) = 0$, we say that there is no correlation between the two variables x and y *i.e.*, the two variables are uncorrected.
   (*iv*) If $\rho(X, Y) > 0$, we say that the correlation between x and y is positive (direct).
   (*v*) If $\rho(X, Y) < 0$, we say that the correlation between x and y is negative (indirect).
2. It is independent of the change of origin and scale.
3. It is pure number and hence free from any of the units.
4. It is rigidly defined and hence free from human bias.

Limitation: It is difficult to compute.

SOLVED EXAMPLES

Example 4.76. Find the coefficient of correlation from the following points of observation (1, 3), (2, 2), (3, 5), (4, 4), (5, 6).

Solution.

x_i	y_i	$x_i - \bar{x} = x_i - 3$	$y_i - \bar{y} = y_i - 4$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	3	–2	–1	4	1	2
2	2	–1	–2	1	4	2
3	5	0	1	0	1	0
4	4	1	0	1	0	0
5	6	2	2	4	4	4
$\Sigma x_i = 15$	$\Sigma y_i = 20$	$\Sigma(x_i - \bar{x}) = 0$	$\Sigma(y_i - \bar{y}) = 0$	$\Sigma(x_i - \bar{x})^2 = 10$	$\Sigma(y_i - \bar{y})^2 = 10$	$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 8$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{15}{5} = 3 \text{ and } \bar{y} = \frac{\sum y_i}{n} = \frac{20}{5} = 4$$

$$\rho(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum (x_i - \bar{x})^2\right]\left[\sum (y_1 - \bar{y})^2\right]}}$$

$$= \frac{8}{\sqrt{10} \times \sqrt{10}} = \frac{8}{10} = 0.8.$$

Example 4.77. Calculate the coefficient of correlation between the age of husband and wife from the following data:

Age of husband	35	34	40	43	56	20	38
Age of wife	32	30	31	32	53	20	33

Solution. Let the age of husband by x and the age of wife by y.

x_i	$x - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
35	−3	9	32	−1	1	3
34	−4	16	30	−3	9	12
40	2	4	31	−2	4	−4
43	5	25	32	−1	1	−5
56	18	324	53	20	400	360
20	−18	324	20	−13	169	234
38	0	0	33	0	0	0
266	0	702	231	0	584	600

$$\sum x_i = 266, \sum (x_i - \bar{x}) = 0, \sum (x_i - x)^2 = 702$$

$$\sum y_i = 231, \sum (y_i - \bar{y}) = 0, \sum (y_i - \bar{y})^2 = 584$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 600$$

$$\therefore \quad \rho(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})\right)^2 \times \sum (y_i - \bar{y})^2}} = \frac{600}{\sqrt{702 \times 584}}$$

$$= \frac{600}{640.287} = 0.937.$$

Example 4.78. Calculate the coefficient of correlation between x and y for the following data:

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Solution.

x_i	y_i	$u_i = x_i - 68$	$v_i = y_i - 69$	$u_i v_i$	u_i^2	v_i^2
65	67	− 3	− 2	6	9	4
66	68	− 2	− 1	2	4	1
67	65	− 1	− 4	4	1	16
67	68	− 1	− 1	1	1	1
68	72	0	3	0	0	9
69	72	1	3	3	1	9
70	69	2	0	0	4	0
72	71	4	2	8	16	4
544	552	0	0	24	36	44

$$n = 8, \sum x_i = 544, \sum y_i = 552$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{554}{8} = 68 \text{ (whole number)}$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{552}{8} = 69 \text{ (whole number)}$$

∴ We take, $u_i = x_i - 68$ and $v_i = y_i = -69$.

$$\rho(X, Y) = \rho(U, V) = \frac{n\sum u_i v_i - (\sum u_i)(\sum v_i)}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}}$$

$$= \frac{8\times 24 - 0\times 0}{\sqrt{8\times 36 - 0}\sqrt{8\times 44 - 0}} = \frac{8\times 24}{8\times 6\times 2\sqrt{11}}$$

$$= \frac{2}{\sqrt{11}} = \frac{2}{11}\times\sqrt{11} = \frac{2\times 3.32}{11} = \frac{6.64}{11} = 0.604$$

Example 4.79. From the following data, compute the coefficient of correlation between x and y.

(*i*) Arithmetic mean of x series is 25 and that of y is 18.

(*ii*) Sum of the products of deviations of x and y series from their repetitive means = 122.

(*iii*) Sum of the squares of deviation from their respective means are 136, 138 respectively for x series and y series.

(*iv*) Number of pairs of values = 15.

Solution. We are given the following information:

$$n = 15, \ \bar{x} = 25, \ \bar{y} = 18$$

$$\sum (x - \bar{x})^2 = 136, \ \sum (y - \bar{y})^2 = 138 \ \sum (x - \bar{x})(y - \bar{y}) = 122$$

Now
$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{122}{\sqrt{136\times 138}} = \frac{122}{136.996} = 0.89.$$

Hence, the coefficient of correlation is 0.89.

Example 4.80. From the data given below, find the number of items n.

$$r = 0.5,\ \sum xy = 120,\ \sigma_y = 8,\ \sum x^2 = 90$$

where x and y are deviations from arithmetic mean.

Solution. Here x and y are deviations from arithmetic mean, *i.e.*, $x = X - \bar{X}$

Now $$\sigma_x = \sqrt{\frac{1}{n}\sum (Y-\bar{Y})^2} = 8 \Rightarrow \sum (Y-\bar{Y})^2 = 64n.$$

Also $$\sum x^2 = \sum (X-\bar{X})^2 = 90.$$

Again $$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}} = 0.5 \Rightarrow \frac{120}{\sqrt{90\times 64n}}$$

$\Rightarrow$ $$0.25 = \frac{120\times 120}{90\times 64n} \Rightarrow 64n = \frac{120\times 120}{90\times 0.25}$$

or $$n = \frac{120\times 120}{64\times 90\times 0.25} = 10.$$

Example 4.81. A computer operator while calculating the correlation coefficient between two variable x and y from 25 pairs of observations obtained the following constants:

$$\Sigma x = 125,\ \Sigma x^2 = 650,\ \Sigma y = 100,\ \Sigma y^2 = 460,\ \Sigma xy = 504.$$

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

Obtain the correct value of the correlation coefficient.

Solution. Here Corrected Σx = Incorrect $\Sigma x - (6 + 8) + (8 + 6) = 125$

Corrected Σy = Incorrect $\Sigma y - (14 + 6) + (12 + 8) = 100$

Corrected $\Sigma x^2 = 650 - (6^2 + 8^2) + (8^2 + 6^2) = 650$

Corrected $\Sigma y^2 = 460 - (14^2 + 6^2) + (12^2 + 8^2) = 436$

Corrected $\Sigma xy = 508 - (84 + 48) + (96 + 48) = 520$

Correct value of correlation coefficient is:

$$r = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sqrt{\left\{\sum x^2 - \frac{1}{n}(\sum x)^2\right\}\left\{\sum y^2 - \frac{1}{n}(\sum y)^2\right\}}}$$

$$= \frac{520 - \frac{1}{25}\times 125\times 100}{\sqrt{\left\{650 - \frac{125\times 125}{25}\right\}\left\{436 - \frac{100\times 100}{25}\right\}}}$$

$$= \frac{20}{\sqrt{25\times 36}} = \frac{20}{5\times 6} = \frac{4}{6} = \frac{2}{3} = 0.67.$$

Example 4.82. Establish the formula

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\,\sigma_y}$$

where r is the coefficient of correlation between x and y and σ_x, σ_y, σ_{x-y} are concerned standard deviations. Hence, evaluate r from the following data:

x	21	23	30	54	57	58	72	78	87	90
y	60	71	72	83	110	84	100	92	113	135

Solution. Let $z = x - y$

$\therefore$ $$\bar{z} = \bar{x} - \bar{y}$$

$\therefore$ $$z - \bar{z} = (x - \bar{x}) - (y - \bar{y})$$

or, $$(z - \bar{z})^2 = (x - \bar{x})^2 + (y - \bar{y})^2 - 2(x - \bar{x})(y - \bar{y})$$

Summing up for n terms,

$$\sum (z - \bar{z})^2 = \sum (x - \bar{x})^2 + \sum (y - \bar{y})^2 - 2\sum (x - \bar{x})(y - \bar{y})$$

or, $$\frac{\sum (z - \bar{z})^2}{n} = \frac{\sum (x - \bar{x})^2}{n} + \frac{\sum (y - \bar{y})^2}{n} - \frac{2\sum (x - \bar{x})(y - \bar{y})}{n}$$

$\Rightarrow$ $$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y \text{ where } r = \frac{\sum (x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}$$

$\Rightarrow$ $$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y \qquad ...(i)$$

or, $$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y} \qquad ...(ii)$$

Now we shall proceed to obtain r from the given data. Here, $n = 10$,

$$\bar{x} = \frac{\sum x}{n} = \frac{21 + 23 + 30 + 54 + 57 + 58 + 72 + 78 + 87 + 90}{10} = \frac{570}{10} = 57$$

$$\bar{y} = \frac{\sum y}{n} = \frac{60 + 71 + 72 + 83 + 110 + 84 + 100 + 92 + 113 + 135}{10} = \frac{920}{10} = 92$$

Now we form the table as

x	y	$z = y - y$	$x - \bar{x}$	$y - \bar{y}$	$z - \bar{z}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(z - \bar{z})^2$
21	60	– 39	– 36	– 32	– 4	1296	1024	16
23	71	– 48	– 34	– 21	– 13	1156	441	169
30	72	– 42	– 27	– 20	– 7	729	400	49
54	83	– 29	– 3	– 9	6	9	81	36
57	110	–53	0	18	–18	0	324	324
58	84	–26	1	– 8	9	1	64	81
72	100	–28	15	8	7	225	64	49
78	92	–14	21	0	21	441	0	441
87	113	–26	30	21	9	900	441	81
90	135	–45	33	43	–10	1089	1849	100
		$\Sigma z = -350$				$\Sigma(x - \bar{x})^2 = 5846$	$\Sigma(y - \bar{y})^2 = 4688$	$\Sigma(z - \bar{z})^2 = 1346$

where, $$\bar{z} = \frac{\sum z}{n} = \frac{-350}{10} = -35$$

Now, $$\sigma_x^2 = \frac{1}{n}\sum (x - \bar{x})^2 = \frac{1}{10}(5846) = 584.6$$

$$\sigma_y^2 = \frac{1}{n}\sum (y - \bar{y})^2 = \frac{1}{10}(4688) = 468.8$$

$$\sigma_{x-y}^2 = \sigma_z^2 = \frac{1}{n}\sum (z - \bar{z})^2 = \frac{1}{10}(1346) = 134.6$$

Substituting the values in the formula (*ii*), we get

$$r = \frac{584.6 + 468.8 - 134.6}{2\sqrt{584.6}\sqrt{468.8}} = \frac{918.8}{2(523.50786)}$$

$\Rightarrow$ $$r = 0.8775.$$

Example 4.83. If $z = ax + by$ and r is the correlation co-efficient between x and y, show that $\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y$.

Solution. $z = ax + by$.

$\Rightarrow$ $$\bar{z} = a\bar{x} + b\bar{y},\ z_i = ax_i + by_i$$

$$z_i - \bar{z} = a(x_i - \bar{x}) + b(y_i - \bar{y})$$

Now $$\sigma_z^2 = \frac{1}{n}\sum (z_i - \bar{z})^2 = \frac{1}{n}\sum [a(x_i - \bar{x}) + b(y_i - \bar{y})]^2$$

$$= \frac{1}{n}\sum [a^2 (x_i - \bar{x})^2 + b^2 (y_i - \bar{y})^2 + 2ab(x_i - \bar{x})(y_i - \bar{y})]$$

$$= a^2 \cdot \frac{1}{n}\sum (x_i - \bar{x})^2 + b^2 \cdot \frac{1}{n}\sum (y_i - \bar{y})^2 + 2ab \cdot \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y \qquad \left[\because r = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x\sigma_y}\right]$$

Example 4.84. The Personnel manager of a large chain of retail stores is interseted in studying the relationship between the experience of his sales personnel and their sales performance. For this purpose, he took a sample of 5 sales personnel at random and recorded data regarding their experience in years and sales in ₹ lakhs as follows:

Sales personnel	*Experience (in yrs.)*	*Sales performance (₹ lakhs)*
1	2	20
2	4	12
3	6	18
4	8	10
5	10	40

Calculate Karl Pearson's coefficient of correlation between the experience and sales performance of the sales personnel.

Solution. Let the experience be denoted by x and sales performance be denoted by y. Let us form the table as:

x	y	xy	x^2	y^2
2	20	40	4	400
4	12	48	16	144
6	18	108	36	324
8	10	80	64	100
10	40	400	100	1600
$\Sigma x = 30$	$\Sigma y = 100$	$\Sigma xy = 676$	$\Sigma x^2 = 220$	$\Sigma y^2 = 2568$

Here, $n = 5$

Karl Pearson's coefficient of correlation is given by

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

$$= \frac{(5\times 676) - (30\times 100)}{\sqrt{(5\times 220) - (30)^2}\sqrt{(5\times 2568) - (100)^2}}$$

$$= \frac{3380 - 3000}{\sqrt{1100 - 900}\sqrt{12840 - 10000}} = \frac{380}{\sqrt{200}\sqrt{2840}} = 0.5042$$

Merits and Demerits of Karl Pearson's Coefficient of Correlation

Merits:

1. It is an ideal method of calculating coefficient of correlation.
2. Exact measurement of correlation is possible.
3. It indicates the degree and direction of correlation.
4. It summaries the relationship in one figure.

Demerits:

1. It always assumes that there is a linear relationship between the variable, whether such relationship exists or not.
2. It is not easy to interpret the significance of correlation coefficient. Very often it is misinterpreted.
3. It is unduly affected by the extreme values.
4. As compared to other methods, it is more time consuming.

4.21.8.3. Spearman's Rank Correlation Coefficient Method

Karl Pearson's coefficient of correlation discussed degree of covariability of liner relationship between two quantitative variables. But often we come across situations when definite measurements on the variables are not possible *i.e.*, the factor, under study cannot be measured in quantitative terms. For example, the evaluation of a group of students on the basis of leadership ability, the ordering of women in a beauty contest, the ranking of students by two or more judges in an interview and so on. In all such cases, objects or individuals may be ranked and arranged in order of merit or proficiency on two variable and when these two sets of ranks covery or have agreement between them, we measure the degree of covariation or relationship by **coefficient of rank correlation.**

In rank correlation we may have three types of problems:

(I) when ranks are given.

(II) when ranks are not given.

(III) when equal ranks are given to more than two attributes.

(I) When Ranks are Given

Where actual ranks are given to us the steps required fro computing rank correlation are:

Step 1. Take the difference of two ranks, *i.e.*, $(R_1 - R_2)$ and denote these differences by D.

Step 2. Square these difference and obtain the total $\sum D^2$.

Step 3. Apply the formula: $R = 1 - \dfrac{6\sum D^2}{n(n^2-1)}$

where n is the total number of pairs of observations.

SOLVED EXAMPLES

Example 4.85. Two ladies were asked to rank 7 different types of lipsticks. The ranks given by them are given below:

Lipsticks	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Anita	2	1	4	3	5	7	6
Sunita	1	3	2	4	5	6	7

Calculate Spearman's rank correlation coefficient.

Solution. Let ranks given by Anita be denoted by R_1 and Sunita by R_2.

Calculation of rank correlationcoefficient

R_1	R_2	$D = R_1 - R_2$	D^2
2	1	+1	1
1	3	−2	4
4	2	+2	4
3	4	−1	1
5	5	0	0
7	6	+1	1
6	7	−1	1
			$\Sigma D^2 = 12$

Now $$R = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

Here $\Sigma D^2 = 12, n = 7$

$\therefore$ $$R = 1 - \frac{6\times 12}{7(7^2-1)} = 1 - 0.214 = 0.786.$$

Example 4.86. Ten competitors in a beauty contest are ranked by three judges in the following orders:

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

Use the correlation coefficient to determine which pair of judges has the nearest approach to common taste in beauty.

Solution. Let R_1, R_2, R_3, respectively be the ranks given by first, second and third judge. Let r_{ij} be rank correlation coefficient between the rank given by ith and jth judge, $i \neq j$, $i = 1, 2, 3$, $j = 1, 2, 3$. Let $D_{ij} = R_i - R_j$ be the difference of ranks of an individual given by ith and jth judge.

Computation of rank correlation coefficient

R_1	R_2	R_3	$D_{12}=R_1-R_2$	$D_{13}=R_1-R_3$	$D_{23}=R_2-R_3$	D_{12}^2	D_{13}^2	D_{23}^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	8	1	2	1	1	4	1
8	9	7	−1	1	2	1	1	4
			$\Sigma D_{12}=0$	$\Sigma D_{13}=0$	$\Sigma D_{23}=0$	$\Sigma D_{12}^2=200$	$\Sigma D_{13}^2=60$	$\Sigma D_{23}^2=214$

Here $n = 10$

$$R_{12} = 1-\frac{6\sum D_{12}^2}{n(n^2-1)} = 1-\frac{6\times 200}{10\times 99} = -\frac{7}{33} = -0.2121$$

$$R_{13} = 1-\frac{6\sum D_{13}^2}{n(n^2-1)} = 1-\frac{6\times 60}{10\times 99} = -\frac{7}{11} = 0.6363$$

$$R_{23} = 1-\frac{6\sum D_{23}^2}{n(n^2-1)} = 1-\frac{6\times 214}{10\times 99} = -\frac{-49}{165} = 0.2970$$

Since R_{13} is maximum so the pair of first and third judge has the nearest approach to the common taste of beauty.

Example 4.87. The coefficient of rank correlation of marks obtained by 10 students in English and Economics was found to be 0.5. If was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

Solution. $$R = 1-\frac{6\sum D^2}{n(n^2-1)}$$

Substituting the value in the above formula, we get

$$0.5 = 1-\frac{6\sum D^2}{10(100-1)} \Rightarrow \frac{6\sum D^2}{990} = 1-0.5 = 0.5$$

$$\therefore \quad 6\sum D^2 = 0.5\times 990 \Rightarrow \sum D^2 = \frac{0.5\times 990}{6} = 82.5$$

Correct value of $$\sum D^2 = 82.5 - 3^2 + 7^2 = 122.5$$

Correct $$R = 1-\frac{6\times 122.5}{10\times 99} = 1-\frac{735}{990} = \frac{225}{990} = 0.2676.$$

Example 4.88. The coefficient of rank correlation between the debenture prices and share prices of a company was 0.8. If the sum of squares of the difference in ranks was 33, find the value of n.

Solution.

$$R = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

Here $R = 0.8$, $\sum D^2 = 33$

From (i), $0.8 = 1 - \frac{6\times 33}{n(n^2-1)}$

$$\Rightarrow \quad n(n^2-1) = \frac{198}{0.2} = 990 \Rightarrow n^2 - n = 990$$

$$\Rightarrow \quad n^3 - 1000 - n + 10 = 0 \quad \Rightarrow \quad n^3 - (10)^3 - (n-10) = 0$$

$$\Rightarrow \quad (n-10)(n^2 + 10n + 100 - 1) \quad \Rightarrow \quad (n-10)(n^2 + 10n + 99) = 0$$

$$\Rightarrow \quad n = 10$$

Hence, number of item in the group is 10. ($\because$ n cannot be imaginary)

(II) When Ranks are Not Given

In this case we are given only the data. We assign ranks to both the series x and y by giving the rank 1 to highest values in both the series (or to the lowest values in both the series) and so on.

SOLVED EXAMPLES

Example 4.89. From the data given below, calculate the coefficient of rank correlation between X and Y.

X	78	89	97	69	59	79	68	57
Y	125	137	156	112	107	136	123	108

Solution. Here $n = 8$

Computation of rank correlation coefficient

X	Y	Rank in X (R_1)	Rank in Y (R_2)	Rank difference $D = R_1 - R_2$	D^2
78	125	4	4	0	0
89	137	2	2	0	0
97	156	1	1	0	0
69	112	5	6	−1	1
59	107	7	8	−1	1
79	136	3	3	0	0
68	123	6	5	1	1
57	108	8	7	1	1
				$\Sigma D = 0$	$\Sigma D^2 = 4$

Coefficients of rank correlation:

$$R = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\times 4}{8(64-1)} = 1 - \frac{3}{63} = 0.952.$$

Example 4.90. A random sample of 5 college students is selected and their grades in Mathematics and Statistics are found to be:

	1	2	3	4	5
Mathematics	85	60	73	40	90
Statistics	93	75	65	50	80

Calculate Spearman's rank correlation coefficient.

Solution. *Computation of rank correlation coefficient*

Marks in Mathematics X_i	*Rank in X* R_1	*Marks in Statistics* Y_i	*Rank in Y* R_2	*Rank difference* $D = R_1 - R_2$	D^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	65	4	−1	1
40	5	50	5	0	0
90	1	80	2	−1	1
					$\Sigma D^2 = 4$

$$\text{Spearman's rank correlation} = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\times 4}{5(5^2-1)}$$

$$= 1 - \frac{24}{5(25-1)} = 1 - \frac{24}{5\times 24} = 1 - \frac{1}{5} = \frac{4}{5} = 0.8.$$

(III) When Equal Ranks are Given to More than Two Attributes

It two or more individuals are placed together in any classification with respect to an attribute, *i.e.*, if in case of variable data, their are more than one item with the same rank in either or both the series (*i.e.*, Tie rank), then the Spearman's Rank Correlation coefficient formula given above does not give the correlation coefficient for Tie Rank. The problem is solved by assigning average rank to each of these individuals who are put in tie. For example, suppose an item is repeated at rank 5th (*i.e.*, the 5th and 6th items are having same values), then the common rank assigned to 5th and 6th items is $(5 + 6)/2 = 5.5$, which is the average of 5 and 6, the ranks which these items would have been assigned if they were different. The next rank assigned will be 7. But if an item is repeated thrice at rank 2, then the common rank assigned to each value will be $(2 + 3 + 4)/3 = 3$, which is the arithmetic mean of 2, 3 and 4. The next rank to be assigned would be 5. In order to find the Rank Correlation coefficient of Repeated Ranks or Tie Ranks, an Adjustment or Correction Factor is added to the Spearman's Rank Correlation formula, *i.e.*, as

$$R = 1 - \frac{6\sum D^2}{n(n^2-1)} \qquad \ldots(i)$$

Correction Factor. "In the formula (*i*), 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 series."

The modified formula for Tie Rank Correlation Coefficient is given by

$$R = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3) + \ldots\ldots\right]}{n(n^2-1)}$$

where $m_1, m_2, m_3, \ldots\ldots$, are the number of times a value is repeated.

SOLVED EXAMPLES

Example 4.91. From the following data of the makes obtained by 8 students in Mathematics and Physics paper compute rank coefficients of correlation.

Marks in Mathematics	15	20	28	12	40	60	20	80
Marks in Physics	40	30	50	30	20	10	30	60

Computation of rank correlation

Marks in Mathematics X	*Rank assigned R_1*	*Marks in Physics Y*	*Rank assigned R_2*	D^2
15	2	40	6	16.00
20	3.5	30	4	0.25
28	5	50	7	4.00
12	1	30	4	9.00
40	6	20	2	16.00
60	7	10	1	36.00
20	3.5	30	4	0.25
80	8	60	8	0.00
				$\Sigma D^2 = 81.5$

$$R = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n^3 - n}$$

Here $\Sigma D^2 = 81.5, n = 8$

The item 20 is repeated 2 times in series X. So $m_1 = 2$. In series Y the item 30 occurs 3 times and so $m_2 = 3$.

Substituting these values in the above formula.

$$R = 1 - \frac{6\left[81.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right]}{8^3 - 8}$$

$$= 1 - \frac{6(81.5 + 0.5 + 2)}{504} = 1 - \frac{6 \times 84}{504} = 1 - \frac{504}{504} = 0.$$

Example 4.92. From the following table, calculate the rank correlation coefficient:

X	48	33	40	9	16	16	65	24	16	57
Y	13	13	24	6	15	4	20	9	6	19

Solution. We prepare the following table.

Computation of rank correlation coefficient

X	Y	Rank X R_1	Rank Y R_2	$D = R_1 - R_2$	D^2
48	13	8	5.5	2.5	6.25
33	13	6	5.5	0.5	0.25
40	24	7	10	−3	9
9	6	1	2.5	−1.5	2.25
16	15	3	7	−4	16
16	4	3	1	2	4
65	20	10	9	1	1
24	9	5	4	1	1
16	6	3	2.5	0.5	0.25
57	19	9	8	1	1
					$\Sigma D^2 = 41$

In the X-series, we notice that the value 16 is repeated thrice. The common rank given to these values is 3, which is the average of 2, 3 and 4 $\left[i.e., \frac{(2+3+4)}{3} = 3\right]$, the ranks which these values would have assumed if they were different. The next rank given to the next value is 5. Here $m_1 = 3$. The correction factor for X-series is.

$$\frac{m_1(m_1^2 - 1)}{12} = \frac{3(3^2 - 1)}{12} = 2.$$

In the Y-series, the value 6 and 13 are repeated twice. The value of 6 occurs twice and its average rank is $2.5 = \frac{(2+3)}{2}$. The next value 9 is assigned the rank 4. Here $m_2 = 2$. Again the value 13 is

repeated twice at different places. The average rank is $5.5 = \frac{5+6}{2}$. The next value 5 is assigned the rank 7. Here $m_3 = 2$. The total correction for the Y-series is

$$\frac{m_2\left(m_2^2-1\right)}{12}+\frac{m_3\left(m_3^2-1\right)}{12}=\frac{2(2^2-1)}{12}+\frac{2(2^2-1)}{12}=0.5+0.5=1.$$

Thus the rank correlation is:

$$R=1-\frac{6\left[\sum D^2+\frac{m_1\left(m_1^2-1\right)}{12}+\frac{m_2\left(m_2^2-1\right)}{12}+\frac{m_3\left(m_3^2-1\right)}{12}\right]}{n(n^2-1)}$$

$$=1-\frac{6[41+2+0.5+0.5]}{10\,(10^2-1)}=1-\frac{264}{990}=1-0.267=0.733.$$

Merits and Demerits of Rank Correlation Coefficient Method

Merits:

1. It is simple to calculate and easy to understand.
2. This method is very useful where the data are of a qualitative nature.
3. When we are given the ranks instead of actual data, this is the only method for finding out the degree of correlation.

Demerits:

1. This method cannot be used for calculating correlation in a grouped frequency distributions. It is applicable only to individual observations.
2. If the number of values is quite large, it becomes a difficult task to ascertain the ranks and their differences.
3. This method lacks precision as compared to Karl Pearson's method. It uses ranks only. Original values are not taken into account.

4.21.9. Calculation of Co-efficient of Correlation for a Bivariate Frequency Distribution

If the diverted data on x and y is presented on a two way correlation table and f is the frequency of a particular rectangle in the correlation table, then

$$r_{xy}=\frac{\sum fxy-\frac{1}{n}\sum fx\sum fy}{\sqrt{\left[\sum fx^2-\frac{1}{n}(\sum fx)^2\right]\left[\sum fy^2-\frac{1}{n}(\sum fy)^2\right]}}$$

Since change of origin and scale do not affect the co-efficient of correlation.

$r_{xy} = r_{uv}$ where the new variables u, and properly chosen.

4.22. COEFFICIENT OF DETERMINATION

We have seen that correlation indicates the amount of variation of one variable which is associated with or which is accounted for by the variation in another variable. A more easily understood and in

certain cases a better measure to fulfill this purpose is the *coefficient of determination*. It indicates the actual percentage of the portion of one variable which is associated with the other or the percentage variation in one variable which is accounted for by the other. Coefficient of determination is the square of the coefficient of correlation. The relationship between the coefficient of correlation and coefficient of determination is of the following type:

$r\%$	$n\%$	$r\%$	$n\%$
1.00	0.40	0.16	0.16
0.90	0.81	0.30	0.09
0.80	0.64	0.20	0.04
0.60	0.36	0.10	0.01
0.50	0.25		

The coefficient of determination is a better measure than the coefficient of correlation. If we compare the two coefficients of correlation one of which is + 0.6 and the other + 0.3, we shall have the impression that the correlation in the first case in twice as high as in the second but the truth is that the correlation in the first case in four times as high as in the second case. This fact is clearly indicated by the coefficient of determination. The coefficient of determination in these cases would be respectively + 0.36 + 0.09. If the coefficient of determination is + 0.81 it means that 81% of the variations in the relative series are due to variations in the subject series and the remaining 19% due to other factors. In case the coefficient of correlation is + 0.9, we cannot say that 90% of the variations of the relative series are due to the variations in the subject series. We shall have to find out the square of the coefficient of correlation to find out this percentage. As such sometimes the coefficient of correlation may actually give misleading conclusions. In coefficient of determination there is no such confusion.

SOLVED EXAMPLES

Example 4.93. The following table gives according to age the frequency of marks obtained by 100 students in an intelligence test:

Marks \ Age (in years)	18	19	20	21	Total
10–20	4	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 coefficient of correlation between age and intelligence.

Solution. Let age and intelligence be denoted by x and y respectively.

Mid value	x / y	18	19	20	21	f	u	fu	fu^2	fuv
15	10–20	4	2	2		8	–3	–24	72	80
25	20–30	5	4	6	4	19	–2	–38	76	20
35	30–40	6	8	10	11	35	–1	–35	35	9
45	40–50	4	4	6	8	22	0	0	0	0
55	50–60		2	4	4	10	1	10	10	2
65	60–70		2	3	1	6	2	12	24	–2
	f	19	22	31	28	100	Total	–75	217	59
	v	–2	–1	0	1	Total				
	fv	–38	–22	0	28	–32				
	fv^2	76	22	0	28	126				
	fuv	56	16	0	–13	59				

Let us define two new variables u and v as $u = \frac{y-45}{10}$, $v = x - 20$

$$r_{xy} = \frac{\sum fuv - \frac{1}{n}\sum fu \sum fv}{\sqrt{\left[\sum fu^2 - \frac{1}{n}(\sum fu)^2\right]\left[\sum fv^2 - \frac{1}{n}(\sum fv)^2\right]}}$$

$$= \frac{59 - \frac{1}{100}(-75)(-32)}{\sqrt{\left[217 - \frac{1}{100}(-75)^2\right]\left[126 - \frac{1}{100}(-32)^2\right]}} = \frac{59 - 24}{\sqrt{\frac{643}{4} \times \frac{2894}{25}}} = 0.25.$$

Example 4.94. Calculate the coefficient of correlation from the following bivariate frequency distribution:

Sales revenue (₹ in lakh)	Advertising expenditure (₹ in '000)				Total
	5–10	10–15	15–20	20–25	
75–125	4	1	—	—	5
125–175	7	6	2	1	16
175–225	1	3	4	2	10
225–275	1	1	3	4	9
Total	13	11	9	7	40

Solution. Let advertising expenditure and sales revenue be represented by variables x and y, respectively. The calculations for correlation coefficient are shown below:

Revenue (y)	Mid-value (m)	$x \rightarrow$ Mid-value (m) / d_x / d_y	Advertising expenditure 5–10, 7.5, −1	10–15, 12.5, 0	15–20, 17.5, 1	20–25, 22.5, 2	Total, f	fd_y	fd_y^2	fd_xd_y
75–125	100	−2	4 (8)	1 (0)	— (0)	— (0)	5	−10	20	8
125–175	150	−1	7 (7)	6 (0)	2 (−2)	1 (−2)	16	−16	16	3
175–225	200	0	1 (0)	3 (0)	4 (0)	2 (0)	10	0	0	0
225–275	250	1	1 (−1)	1 (0)	3 (3)	4 (8)	9	9	9	10
		Total, f	13	11	9	7	$n = 40$	$\Sigma d_y = -17$	$\Sigma dy^2 = 45$	$\Sigma fd_xd_y = 21$
		fd_x	−13	0	9	14	$\Sigma fd_x = -10$			
		fd_x^d	13	0	9	28	$\Sigma fd_x^2 = 50$			
		fd_xd_y	14	0	1	6	$\Sigma fd_xd_y = 21$			

where, $d_x = (m - 12.5)/5$ and $dy = (m - 200)/50$

Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$r = \frac{n\sum fd_xd_y - (\sum fd_x)(\sum fd_y)}{\sqrt{n\sum fd_x^2 - (\sum fd_x)^2}\sqrt{n\sum fd_y^2 - (\sum fd_y)^2}}$$

$$= \frac{40\times 21\times 10\times -17}{\sqrt{40\times 50 - (10)^2}\sqrt{40\times 45 - (-17)^2}}$$

$$= \frac{840 + 170}{\sqrt{1900}\sqrt{1511}} = \frac{1010}{1694.373} = 0.596$$

Interpretation: Since the value of r is positive, advertising expenditure and sales revenue are positively correlated to the extent of 0.596. Hence, we conclude that as expenditure on advertising increases, the sales revenue also increases.

EXERCISE 4.1

1. Which in your opinion is an ideal measure of correlation? Give reasons in support of your answer.
2. What are the properties of coefficient of correlation?
3. What are the merits and demerits of Karl Pearson's coefficient of correlation?

4. Point out the usefulness of 'rank difference method'.
5. Draw a scatter diagram from the following data:

Height (inches)	62	72	70	60	67	70	64	65	60	70
Weights (kg)	50	65	63	52	66	60	59	58	54	65

6. Following are the heights and weights of 10 students of a class, draw a scatter diagram:

Heights (inches)	62	72	68	58	65	70	66	63	60	72
Weights (kg)	50	65	63	50	64	60	61	55	54	55

7. From the following table calculate the coefficient of correlation by Karl Pearson's method:

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

Arithmetic mean of X and Y series are 6 and 8 respectively. **[Ans. $r = 0.82$]**

8. Calculate the correlation X and Y from the following data:

X	61	72	73	63	84	80	66	76
Y	40	52	49	43	61	58	42	58

[Ans. $r = +0.97$]

9. Find the coefficient of correlation between the values of x and y.

x	1	3	5	7	8	10
y	8	12	15	17	18	20

[Ans. $r = 0.9879$]

10. From the data given below, compute Karl Pearson's coefficient of correlation:

	X-series	Y-series
Number of item	15	15
Arithmetic mean	25	18
Squares of deviations from arithmetic mean	136	138

Summation of products of deviations from the arithmetic means of X and $Y = 122$.

[Ans. $r = +0.89$]

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

Husband's age	24	27	28	28	29	30	32	33	35	35	40
Wife's age	18	20	222	25	22	28	28	30	27	30	22

[Ans. $r = +0.504$]

12. The following data regarding the heights (y) and weight (x) of 100 college students are given:

$$\sum x = 15000,\ \sum x^2 = 2272500,\ \sum y = 6800,$$

$$\sum y^2 = 463025 \text{ and } \sum xy = 1022250$$

Find the correlation coefficient between height and weight. **[Ans.** $r = 0.6$**]**

13. Calculate the rank correlation coefficient for the following data of marks obtained by 8 students in Economics and Statistics.

Students	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Marks in —Economics	15	20	28	12	40	60	20	80
—Statistics	40	30	50	30	20	10	30	60

[Ans. $r = 0$**]**

14. Calculate coefficient of rank correlation from the following data:

Cost (in ₹)	80	78	75	75	68	67	60	59
Scale Price (in ₹)	*12*	13	14	14	14	16	15	17

[Ans. $r = -0.324$**]**

15. The following table gives the number of students having different heights and weights:

Heights in centimeters	***Weight in kilograms***					
	55–60	***60–65***	***65–70***	***70–75***	***75–80***	***Total***
150–155	1	3	7	5	2	18
155–160	2	4	10	7	4	27
160–165	1	5	12	10	7	35
165–170	—	3	8	6	3	20
Total	4	15	37	28	16	100

Find the coefficient of correlation between the heights and weights of the students.
[Ans. 0.0946**]**

16. The following table given the frequency, according to the marks, obtained by 67 students in an intelligence test. Measure the degree of relationship between age and marks:

Test Marks	Age in years				Total
	18	19	20	21	
200—250	4	4	2	1	11
250—300	3	5	4	2	14
300—350	2	6	8	5	21
350—400	1	4	6	10	21
Total	10	19	20	18	67

[Ans. 0.415]

17. Show that if a and b are constants and r is the correlation coefficient between x and y, then 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 $-r$ if they are different.
18. Two independent random variable x and y have the following variances:

$$\sigma_x^2 = 36, \quad \sigma_y^2 = 16$$

Find the coefficient of correlation between $u = x + y$ and $v = x - y$. $\left[\textbf{Ans. } r_{uv} = \frac{5}{13}\right]$

19. x and y are two variates with variances σ_x^2 and σ_y^2 respectively and k is the coefficient of correlation between them. If $u = x + ky$, $v = x + \frac{\sigma_x}{\sigma_y} y$, find the value of k so that u and v are uncorrelated. $\left[\textbf{Ans. } k = \frac{-\sigma_x}{\sigma_y}\right]$
20. Two variates x and y have zero means, the same variance σ^2 and zero correlation coefficient. Show that $u = x \cos \alpha + y$ and α and $v = x \sin a - y \cos a$ have the same variance σ^2 and zero correlation.
21. Write short notes on:
(*a*) Positive and negative correlation, (*b*) Coefficient of determination.

4.23. REGRESSION ANALYSIS

The regression analysis is concerned with the formulation and determination of algebraic expressions for the relationship between the two variables. We use the general form 'regression lines' for these algebraic expressions. These regression lines or the exact algebraic forms of the relationship are then used for predicting the value of on variable from that of the other. Here the variable whose value is to be predicted is called dependent or explained variable and the variable used for prediction is called *independent* or *explanatory variable*.

Remember: By regression we mean the average relationship between two or more variables which can be used for estimating the value of one variable from the given values of one or more variables.

In a bivariate distribution, the analysis is restricted to only two variables only.

Difference between Correlation and Regression Analysis

1. Correlation coefficient is quantitative measure of the extent or degree of linear relationship between the two variables. On the other hand regression means an average relationship between the two variables.
2. Correlation does not necessarily establish cause and effect relationship. However, in regression analysis there is a clear indication of cause and effect relationship. Here, the independent variable is the cause and dependent variable the effect.
3. Correlation may be nonsense or spurious while there is nothing like non sense regression.
4. In correlation analysis, r_{xy} measures the linear relationship between the variable x and y. Here $r_{xy} = r_{yx}$ *i.e.*, it is immaterial which of the two variables is taken as dependent or independent. However, in regression analysis, the identity of variable *i.e.*, which is dependent and which one is independent in important. As will be seen later, the two regression coefficients b_{xy} and b_{yx} are not symmetric like correlation coefficient *i.e.*, $b_{xy} \neq b_{yx}$.
5. Correlation coefficient r_{xy} is a relative measure of the linear relationship between x and y, while the regression coefficients, b_{yx} and b_{xy}, are absolute measures of change in the value of one variables corresponding to a unit change in value of another variable.
6. Whereas correlation analysis is confined only to the study of linear relationship between two variables, the regression analysis deals with linear and non-linear relationship between variables.

4.23.1. Regression Lines

For a simple regression analysis there are two regression lines called X on Y and Y on X, where X and Y are two variables. X on Y lines denotes that X is dependent and Y independent, *i.e.*, X depends on Y. Similarly Y on X denotes that Y depends on X.

Let $\bar{x}$ and $\bar{y}$ be the means for series X and series Y and σ_X the standard deviation for X series and σ_Y for Y-series. Let r be the correlation coefficient between them; then the regression line for

Y on X $\quad \boxed{Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})}$, regression line for Y on X

X on Y, $\quad \boxed{X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})}$, regression line for X on Y.

Here, the value of r lies between -1 to $+1$ and the terms $r\frac{\sigma_Y}{\sigma_X}$ and $r\frac{\sigma_X}{\sigma_Y}$ are called the *regression coefficients* for Y on X and X on Y series. These coefficients are denoted by b_{YX} and b_{XY}. Hence

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{\sum dx\, dy}{\sum x^2} \quad \ldots(i)$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{\sum dx\, dy}{\sum y^2} \quad \ldots(ii)$$

Multiplying (*i*) and (*ii*)

$$b_{XY}\, b_{YX} = r^2$$

$$\boxed{r = \sqrt{b_{XY} b_{YX}}}$$

4.23.2. Line of Regression of *y* on *x*

For getting the line of regression of y on x, we shall consider y as dependant variable and x as independent variable.

Let $y = a + bx$ be the equation of regression line of y on x.

The residual for i^{th} point is $E_i = y_i - a - bx_i$.

Introduce a new quantity U such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad ...(i)$$

According to the *principle of Least squares*, the constants a and b are chosen in such a way that the sum of the squares of residuals is minimum.

Now, the condition for U to be maximum or minimum is given by

$$\frac{\partial U}{\partial a} = 0 \text{ and } \frac{\partial U}{\partial b} = 0$$

From (*i*),

$$\frac{\partial U}{\partial a} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1)$$

$$\frac{\partial U}{\partial a} = 0 \text{ gives } 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

$$\Rightarrow \quad \boxed{\sum y = na + b\sum x} \qquad ...(ii)$$

Also from (*i*)

$$\frac{\partial U}{\partial b} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0$$

$$\Rightarrow \quad \boxed{\sum xy = a\sum x + b\sum x^2} \qquad ...(iii)$$

(*ii*) and (*iii*) are called *normal equations*.

After solving (*ii*) and (*iii*), we can get the value 'a' and 'b',

$$b = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sum x^2 - \frac{1}{n}(\sum x)^2} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \qquad ...(iv)$$

and

$$a = \frac{\sum y}{n} - b\frac{\sum x}{n} = \bar{y} - b\bar{x} \qquad ...(v)$$

(*v*) gives $\bar{y} = a + b\bar{x}$

Hence, $y = a + bx$ line passes through point $(\bar{x}, \bar{y})$.

Putting $a = y - bx$ in equation of line $y = a + bx$, we get

$$\boxed{y - \bar{y} = (x - \bar{x})} \qquad \ldots(vi)$$

(*vi*) is called *regression line* of y on x and 'b' is called the regression co-efficient of y on x and is usually denoted by b_{yx}.

Hence, eqn. (*vi*) can be rewritten as

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

where $\bar{x}$ and $\bar{y}$ are mean values while

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

In (*iii*), shifting the origin to $(\bar{x}, \bar{y})$, we get

$$\Sigma(x - \bar{x})(y - \bar{y}) = a\Sigma(x - \bar{x}) + b\Sigma(x - \bar{x})^2$$

$$\Rightarrow \quad nr\,\sigma_x\sigma_y = a(0) + bn\sigma_x^2$$

$$\Rightarrow \quad b = r\frac{\sigma_y}{\sigma_x}$$

$$\left\{\begin{array}{l} \because\ \Sigma(x - \bar{x}) = 0 \\ \dfrac{1}{n}\Sigma(x - \bar{x})^2 = \sigma_x^2 \\ \text{and } \dfrac{\Sigma(x - \bar{x})(y - \bar{y})}{n\,\sigma_x\sigma_y} = r \end{array}\right\}$$

Hence, regression coefficient b_{yx} can also be defined as

$$\boxed{b_{yx} = r\frac{\sigma_y}{\sigma_x}}$$

where r is the coefficient of correlation while σ_x and σ_y are the standard deviation of x and y series respectively.

4.23.3. Line of Regression of *x* on *y*

Proceeding in the same, we can derive the regression line of x on y as

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

where b_{xy} is the regression coefficient of x on y and is given by

$$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2} \qquad \text{or} \qquad b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

where the terms have their usual meanings.

4.23.4. Properties of Regression Co-efficients

Property I. Correlation co-efficient is the geometric mean between the regression co-efficients.

The co-efficient of regression are $\frac{r\sigma_y}{\sigma_x}$ and $\frac{r\sigma_x}{\sigma_y}$.

Geometric mean between them $= \sqrt{\frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = r =$ co-efficient of correlation.

Property II. If on of the regression co-efficients is greater than unity, the other must be less than unity.

The two regression co-efficients are $b_{yx} = \frac{r\sigma_y}{\sigma_x}$ and $b_{xy} = \frac{r\sigma_x}{\sigma_y}$.

Let $b_{yx} > 1$, then $\frac{1}{b_{yx}} < 1$ ...(*i*)

Since $b_{xy} \cdot b_{xy} = r^2 \leq 1$ ($\because -1 \leq r \leq 1$)

$\therefore$ $b_{xy} \leq \frac{1}{b_{yx}} < 1.$ [using (*i*)]

Similarly, if $b_{xy} > 1$, then $b_{yx} < 1$.

Property III. Arithmetic mean of regression co-efficients is greater than the correlation co-efficient.

i.e., $\frac{b_{yx} + b_{xy}}{2} > r$

or $r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y} > 2r$

or $\sigma_x^2 + \sigma_y^2 > 2\sigma_x\sigma_y$

or $(\sigma_x - \sigma_y)^2 > 0$ which is true.

Property IV. Regression co-efficients are independent of the origin but not of scale.

Proof. Let $u = \frac{x-a}{h}$ and $v = \frac{y-b}{k}$, where a, b, h and k are constants

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = r \cdot \frac{k\sigma_v}{h\sigma_u} = \frac{k}{h}\left(\frac{r\sigma_v}{\sigma_u}\right) = \frac{k}{h} b_{vu}$$

Similarly, $b_{xy} = \frac{h}{k} b_{uv}$.

Thus b_{yx} and b_{xy} are both independent of a and b but not of h and k.

Property V. The correlation co-efficient and the two regression co-efficients have same sign.

Proof. Regression co-efficient of y on $x = b_{yx} = r\frac{\sigma_y}{\sigma_x}$

Regression co-efficient of x on $y = b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Since σ_x and σ_y are both positive; b_{yx}, b_{xy} and r must have be same sign.

4.23.5. Angle between Two Lines of Regression

If θ is the acute angle between the two regression lines in the case of two variables x and y, show that

$$\tan\theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2},$$ where r, σ_x, σ_y have their usual meanings.

Explain the significance of the formula when $r = 0$ and $r = \pm 1$.

Proof. Equations to the lines of regression of y on x and x on y are

$$y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x}) \text{ and } (x - \bar{x}) = \frac{r\sigma_x}{\sigma_y}(y - \bar{y}) \quad \ldots(i)$$

Their slopes are $m_1 = \dfrac{r\sigma_y}{\sigma_x}$ and $m_2 = \dfrac{\sigma_y}{r\sigma_x}$ $\quad \ldots(ii)$

$$\therefore \quad \tan\theta = \pm\frac{m_2 - m_1}{1 + m_2 m_1} = \pm\frac{\dfrac{\sigma_y}{r\sigma_x} - \dfrac{r\sigma_y}{\sigma_x}}{1 + \dfrac{\sigma_y^2}{\sigma_x^2}} \quad \{\text{using } (i) \text{ and } (ii)\}$$

$$= \pm\frac{1 - r^2}{r} \cdot \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} = \pm\frac{1 - r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

Since $r^2 \leq 1$ and σ_x, σ_y are positive

$\therefore$ +ve sign gives the acute angle between the lines.

Hence, $$\tan\theta = \frac{1 - r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

when $r = 0$, $\theta = \dfrac{\pi}{2}$ $\therefore$ The two lines of regression are perpendicular to each other.

Hence, the estimated values of y is the same for all values of x and vice-versa.

When $r = \pm 1$, $\tan\theta = 0$ so that $\theta = 0$ or π.

Hence, the lines of regression coincide and there is perfect correlation between the two variates x and y.

SOLVED EXAMPLES

Example 4.95. Obtain the line of regression of y on x for the data given below:

x	1.53	1.78	2.60	2.95	3.42
y	33.50	36.30	40.00	45.80	53.50

Solution. The line of regression of y on x is given by

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

where b_{yx} is the coefficient of regression given by

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \quad \ldots(ii)$$

Now we form the table as,

x	y	x^2	xy
1.53	33.50	2.3409	51.255
1.78	36.30	2.1684	64.614
2.60	40.00	6.76	104
2.95	45.80	8.7025	135.11
3.42	53.50	11.6964	182.97
$\Sigma x = 12.28$	$\Sigma y = 209.1$	$\Sigma x^2 = 32.6682$	$\Sigma xy = 537.949$

For $x = 5$

$$b_{yx} = \frac{(5\times 537.949) - (12.28\times 209.1)}{(5\times 32.6682) - (12.28)^2} = \frac{121.997}{12.543} = 9.726$$

Also, $$\text{mean } \bar{x} = \frac{\Sigma x}{n} = \frac{12.28}{5} = 2.456$$

and $$y = \frac{\Sigma y}{n} = \frac{20.91}{5} = 41.82$$

$\therefore$ From (i), we get

$$y - 41.82 = 9.726(x - 2.456) = 9.726x - 23.887$$

$$y = 19.932 + 9.726x$$

which is the required line of regression of y on x.

Example 4.96. The following data regarding the heights (y) and weights (x) of 100 college student are given:

$\Sigma x = 15{,}000$, $\Sigma x^2 = 22{,}72{,}500$, $\Sigma y = 6{,}800$, $\Sigma y^2 = 4{,}63{,}025$ and $\Sigma xy = 10{,}222{,}50$.

Find the equation of regression line of height on weight.

Solution. $$\bar{x} = \frac{\Sigma x}{n} = \frac{15000}{100} = 150$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{6800}{100} = 68$$

Regression coefficient of y on x,

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{(100\times 1022250) - (15000\times 6800)}{(100\times 2272500) - (15000)^2}$$

$$= \frac{102225000 - 102000000}{227250000 - 225000000} = \frac{225000}{2250000} = 0.1$$

Regression line of height (y) on weight (x) is given by

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

$\Rightarrow \quad y - 68 = 0.1\,(x - 150)$

$\Rightarrow \quad y = 0.1x - 15 + 68$

$\Rightarrow \quad y = 0.1x + 53.$

Example 4.97. If the regression coefficients are 0.8 and 0.2, what would be the value of co-efficient of correlation?

Solution. We know that $r^2 = b_{yx} \cdot b_{xy} = 0.8 \times 0.2 = 0.16$ $\quad [\because\ b_{yx} = 0.8;\ b_{xy} = 0.2]$

Since r has the same sign as both the regression coefficients b_{yx} and b_{xy}.

Hence, $\quad r = \sqrt{0.16} = 0.4$

Example 4.98. Calculate linear regression co-efficients from the following:

x	1	2	3	4	5	6	7	8
y	3	7	10	12	14	17	20	24

Solution. Linear regression coefficients are given by

$$b_{yx} = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \quad \text{and} \quad b_{xy} = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$

x	y	x^2	y^2	xy
1	3	1	9	3
2	7	4	49	14
3	10	9	100	30
4	12	16	144	48
5	14	25	196	70
6	17	36	289	102
7	20	49	400	140
8	24	64	576	192
$\Sigma x = 36$	$\Sigma y = 107$	$\Sigma x^2 = 204$	$\Sigma y^2 = 1763$	$\Sigma xy = 599$

Here $n = 8$

$$\therefore \quad b_{yx} = \frac{(8\times 599) - (36\times 107)}{(8\times 204) - (36)^2} = \frac{4792 - 3852}{1632 - 1296} = 2.7976$$

and $$b_{xy} = \frac{(8\times 599) - (36\times 107)}{(8\times 1763) - (107)^2} = \frac{940}{2655} = 0.3540$$

Example 4.99. From the given data obtain the two regression equations using the method of least squares.

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

Solution. Computation of regression equation

X	Y	XY	X^2	Y^2
2	5	10	4	25
4	7	28	16	49
6	9	54	36	81
8	8	64	64	64
10	11	110	100	121
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma XY = 266$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$

For Y on X, normal equations [$\because n = 5$]

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

$\Rightarrow$
$$40 = 5a + 30b$$
and
$$266 = 30a + 220b$$

Solving these two equations, we have

$$a = 4.1, b = 0.65$$

So the required equation $Y = a + bX$

$$Y = 4.1 + 0.65X$$

For X on Y, normal equations

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

and

$\Rightarrow$
$$30 = 5a + 40b$$
$$266 = 40a + 340b$$

On solving, $a = -4.4, b = 1.3$

$\therefore$ The required equation is

$$X = a + bY$$
$$X = -4.4 + 1.3Y$$

Example 4.100. The following table gives age (x) in years of cars and annual maintenance cost (y) in hundred rupees:

X	1	3	5	7	9
Y	15	18	21	23	22

Estimate the maintenance cost for a 4 year old car after finding the regression equation.

Solution.

x	y	xy	x^2
1	15	15	1
3	18	54	9
5	21	105	25
7	23	161	49
9	22	198	81
$\Sigma x = 25$	$\Sigma y = 99$	$\Sigma xy = 533$	$\Sigma x^2 = 165$

Here, $n = 5$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{25}{5} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{99}{5} = 19.8$$

$$\therefore \quad b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{(5\times 533) - (25\times 99)}{(5\times 165) - (25)^2} = 0.95$$

Regression line of y on x is given by

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

$$\Rightarrow \quad y - 19.8 = 0.95(x - 5)$$

$$\Rightarrow \quad y = 0.95x + 15.05$$

When $x = 4$ years, $y = (0.95 \times 4) + 15.05 = 18.85$ hundred rupees = ₹ 1885.

Example 4.101. In a partially destroyed laboratory record of an analysis of a correlation data, the following results only are eligible:

Variance of $x = 9$

Regression equations: $8x - 10y + 66 = 0$, $40x - 18y = 214$.

What were: (a) the mean values of x and y, (b) the co-efficient of correlation between x and y, (c) the standard deviation of y and .

Solution. (a) Since both the lines of regression pass through the point $(\bar{x}, \bar{y})$ therefore, we have

$$8\bar{x} - 10\bar{y} + 66 = 0 \quad ...(i)$$

$$40\bar{x} - 18\bar{y} - 214 = 0 \quad ...(ii)$$

Multiplying (i) by 5,

$$40\bar{x} - 50\bar{y} + 330 = 0 \quad ...(iii)$$

Subtracting (iii) from (ii),

$$32\bar{y} - 544 = 0 \qquad \therefore \ \bar{y} = 17$$

$\therefore$ From eqn. (i),

$$8\bar{x} - 170 + 66 = 0 \text{ or } 8\bar{x} = 104 \ \therefore \ \bar{x} = 13$$

(*b*) Variance of $x = \sigma_x^2 = 9$

$\therefore \quad \sigma_x = 3.$

The equations of lines of regression can be written as

$$y = 0.8x + 6.6 \quad \text{and} \quad x = 0.45y + 5.35$$

$\therefore$ The regression coefficient of y on x is $\dfrac{r\sigma_y}{\sigma_x} = 0.8$ ...(*iv*)

The regression coefficient of x on y is $\dfrac{r\sigma_x}{\sigma_y} = 0.45$ ...(*v*)

Multiplying (*iv*) and (*v*), $r^2 = 0.8 \times 0.45 = 0.36 \therefore \quad r = 0.6$

(+ve sign with square root is taken because regression coefficients are +ve)

(*c*) From (*iv*), we have

$$\frac{r\sigma_y}{\sigma_x} = 0.8$$

$$\Rightarrow \quad \sigma_y = \frac{0.8\sigma_x}{r} = \frac{0.8 \times 3}{0.6} = 4.$$

Example 4.102. A study of prices of a certain commodity at Hapur and Kanpur yields the following data:

	Hapur (₹)	*Kanpur* (₹)
Average price/kilo	2.463	2.797
Standard deviation	0.326	0.207

Correlation coefficient between prices at Hapur and Kanpur is 0.774.

Estimate, from the above data, the most likely price at hapur corresponding to the price of ₹ 3.052 per kilo at Kanpur.

Solution. Here $\bar{x} = 2.463, \ \bar{y} = 2.797, \sigma_x = 0.326, \sigma_y = 0.207$

and $r = 0.774$

Regression co-efficient $b_{xy} = r\dfrac{\sigma_x}{\sigma_y} = (0.774)\left(\dfrac{0.326}{0.207}\right) = 1.218956522$

Regression line of x on y is given by

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

$$\Rightarrow \quad x - 2.463 = 1.218956522\,(y - 2.797)$$

$$x = 1.218956522y - 0.9464213$$

when $y = 3.052$, $\quad x = 1.218956522\,(3.052) - 0.9464213 = 2.7738$

Example 4.103. The data for advertising and sales are given below:

	Adv. Exp (X) (₹ *lakhs*)	*Sales (Y)* (₹ *lakhs*)
Mean	10	90
S.D.	3	12