9.10 INTRODUCTION

Here we shall study some of the probability distributions (discrete and continuous) including their expected values, moment generating function. The probability distribution is the outcome of the different probabilities taken by this function of the random variable X.

A random variable can be either discrete or continuous. A random variable is said to be discrete if the set of values defined by it over the sample space is finite, while a random variable is continuous, If it can assume any real value in an internal, and if the random variable X is a discrete one, the probability function P(X) is called **"Probability mass function"** and its distribution, as "discrete probability distribution" while the random variable X is of continuous type, the probability function P(X) is called **"Probability density function"** and its distribution as "continuous probability distribution". Amongst expected frequency distributions, the following distribution we shall discuss:

(1) Binomial Distribution

(2) Poisson Distribution

(3) Normal Distribution.

Discrete Probability Distributions

9.11 BERNOULLI DISTRIBUTION

A random variable X is said to have a Bernoulli distribution with a parameter P if its probability mass function is given by

$$P(X=x)=\begin{cases} P^x(1-P)^{1-x} & ;\ for\ x=0,1 \\ 0 & ;\ otherwise \end{cases}$$

Where the parameter p satisfies $0 \le P \le 1$

A bernoulli trial is random experiment which has two outcomes either a success (S) or a failure (F) occuring with probabilities p and q respectively. If a random variable X defined on this sample space by $X(\text{success})=1, X(\text{failure})=0$, then the probability distribution is given as

X	:	0	1
$P(X)$	:	q	P

9.11.1 Moment's of Bernoulli Distribution

Suppose X be a Benoulli variate, then the r^{th} moment about origin is given as

$$\mu_r' = E\left(X^r\right) = 0^r.q + 1^r\, p = p,\ r = 1,2,3\ .....$$

$$\mu_1' = E[X] = p,\ E\left(X^2\right) = \sum x_i^2\ p_i = 0\times q + 1\times p = p$$

So that variance

$$(X) = \mu_2 = \mu_2' - \left(\mu_1'\right)^2$$

$$\mu_2 = p^2 - p = p(1-p) = pq$$

The manent generating function (m.g.f.) of Bernoulli variate is

$$M_X(t) = e^{0\times t}\ P(X=0) + e^{1\times t}\ P(X=1)$$

$$\boxed{M_X(t) = q + pe^t}$$

9.12 BINOMIAL DISTRIBUTION

Suppose a random experiment be performed repeatedly, each of repitition being called a trial, let there are n independent trials of an experiment, and each trial has two outcomes called success (S) or failure (F) in which the probability of success denoted by P and q is the probability of failure in any trial. It is required to find the probability of getting r successes in n idependent trials, i.e. remaining (n-r) will be failures. If may be shown as follows.

$$\underbrace{p.p.p.p.....p}_{r\ times}\ \underbrace{q.q.q.q.....q}_{(n-r)\ times} = P^r\ q^{n-r}$$

But r success in n trials can occur in $\binom{n}{r}$ or ${}^{n}c_r$ ways and probability for each of these ways is same. Therefore the probability of r successes in n trials in any order is defined as

$$\boxed{P(X=r) = {}^{n}c_r\ p^r\ q^{n-r} \quad ;\ r = 0, 1, 2, ... n;\ p = 1-q}$$

Example : Prove that the sum of probability mass functions for Binomial distribution is one.

Sol. Sum of probability mass function is given by

$$P(X=r) = n_{c_r}\ P^r\ q^{n-r} \qquad ;\ r = 0,1,2,.....n$$

$$= \sum_{r=0}^{n} n_{c_r}\ P^r\ q^{n-r}$$

$$= n_{c_0}\ q^n + n_{c_1}\ pq^{n-1} + n_{c_2}\ p^2 q^{n-2} + ... + n_{c_n}\ p^n$$

$$= (q+p)^n$$

$$\Rightarrow \qquad \sum_{r=0}^{n} P(X=r) = (p+q)^n = 1 \ [\because\ p+q=1]$$

Hence is called the binomial distribution and it is denoted by $B(n, p; r)$, where n and P are known as parameters.

Note :

(1) This property proves that Binomial Distribution is a legitimate probability distribution.

(2) Let as suppose that n-trials constitute an experiment and if this experiment is repeated N-times, then the frequency function of binomial distribution is given by:

$$f(r) = NP(r) = N\ n_{c_r}\ P^r\ q^{n-r},\ r = 0,1,2,.....n$$

9.12.1 Mean and Variance of the Binomial Distribution

Suppose $x_1, x_2, ..., x_n$ are the variate values with corresponding probabilities are $p_1, p_2, p_3, ... p_n$, then

$$\text{Mean}\ (\mu) = E(X) = \sum_{r=0}^{n} x_r\ P(X=r)$$

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

(i) Mean of the Binomial distribution is given by

$$\text{Mean}\ (\mu) = \mu_1' = E(X) = \sum_{r=0}^{n} r\ n_{c_r}\ p^r\ q^{n-r}$$

$$= 0 + 1\ n_{c_1}\ pq^{n-1} + 2\ n_{c_2}\ P^2 q^{n-2} + 3\ n_{c_3}\ P^3\ q^{n-3} + ... + n.p^n$$

$$= npq^{n-1} + n(n-1)\ p^2\ q^{n-2} + \frac{(n-1)(n-2)}{2!}\ P^3 q^{n-3} + ... np^n$$

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

Hence Mean of the Bionamial

$$= np\left[q+p\right]^{n-1} = np1 = np \qquad \left[\because\ p+q=1\right]$$

$$\boxed{\textbf{Mean}\ (\mu) = np}$$

Now $E\left(X^2\right) = \mu_2' = \sum_{r=0}^{n} r^2\ n_{c_r}\ p^r\ q^{n-1}$

$$= \sum_{r=0}^{n} \left[r\ (r-1) + r \right] n_{c_r}\ P^r\ q^{n-r}$$

$$= \sum_{r=0}^{n} r\ (r-1)\, n_{c_r} P^r\, q^{n-r} + \sum_{r=0}^{n} r\ n_{c_r}\ p^r q^{n-r}$$

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

$$\ldots\ldots + n(n-1)\, p^n + \sum_{r=0}^{n} r\, p(X=r)$$

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

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

$$= n(n-1)\ p^2 + np \qquad [\because\ q+p=1]$$

$$\therefore \quad E\left(X^2\right) = \mu_2' = np\left[(n-1)\ p+1\right]$$

$$= np[np-p+1]$$

$$= np\ [np+q]$$

$$\therefore \quad \text{Variance} = E(X^2) - [E(X)]^2 = \mu_2' - (\mu_1')^2$$

$$= n^2P^2 + npq - (np)^2$$

Hence variance of the Binomial distribution is

$$\boxed{\textit{Variance} = npq}$$

and

$$\boxed{\textit{Standard deviation} = \sqrt{\textit{variance}} = \sqrt{npq}}$$

9.12.2 Moments and Moment Generating Function $M_X(t)$ of Binomial Distribution

The moment generating function about the origin is denoted by

$$M_X(t) = E\left[e^{tx}\right] = \sum_{r=0}^{n} e^{tx_r}\ {}^nC_r\ p^r q^{n-r}$$

$$= \sum_{r=0}^{n} {}^nC_r\ P^r q^{n-r}\ e^{tr} = \sum_{r=0}^{\infty} {}^nC_r \left(Pe^t\right)^r\ q^{n-r}$$

$$\therefore \quad \boxed{M_X(t) = \left(q + pe^t\right)^n}$$

Now, we will obtain moments about origin from m.g.f. $\left[M_X(t)\right]$;

$$\mu_1' = \left[\frac{d}{dt} M_X(t)\right]_{t=0} = \left[npe^t \left(q + pe^t\right)^{n-1}\right]_{t=0}$$

$$= np(q + p)$$

$$\therefore \quad \mu_1' = np$$

Which is first moment about origin.

$$\text{Now} \quad \mu_2' = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0} = np\left[\frac{d}{dt}\left(q + pe^t\right)^{n-1} e^t\right]_{t=0}$$

$$= np\left[e^t \left(q + pe^t\right)^{n-1} + (n-1)\, pe^{2t} \left(q + pe^t\right)^{n-2}\right]_{t=0}$$

$$= np\left[1 + (n-1)\, P\right] = np\left[np + 1 - P\right]$$

$$\mu_2' = n^2 p^2 + npq \qquad [\because p+q=1]$$

This is second moment about origin.

and $$\mu_3' = \left[\frac{d^3}{dt^3} M_X(t)\right]_{t=0} = np\left[\frac{d}{dt}\left\{e^t\left(q+pe^t\right)^{n-1}\right\}+(n-1)p\frac{d}{dt}\left\{e^{2t}\left(q+pe^t\right)^{n-2}\right\}\right]_{t=0}$$

$$\mu_3' = np + n(n-1)(n-2)p^3 + 3n(n-1)p^2$$

This represent 3^{rd} moment about origin. Similarly, we can obtain forth moment about the origin is

$$\mu_4' = \left[\frac{d^4}{dt^4} M_X(t)\right]_{t=0} = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3$$

$$+7n(n-1)p^2 + np \quad \text{etc.}$$

9.12.3 Central Moments of Binomial Distribution

The moments about the mean of the Binomial distribution can be obtained by using the following relations, $\mu_1 = 0$ for all distribution.

$$\mu_2 = \mu_2' - \mu_1'^2 = npq$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = npq(q-p)$$

and $$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = npq\left[1+3(n-2)pq\right]$$

9.12.4 Karl Pearson's Coefficients of Skewness and Kurtosis

$\because$ We know that

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{n^2p^2q^2(q-p)^2}{n^3p^3q^3}$$

$$\beta_1 = \frac{(q-p)^2}{npq} = \frac{(1-2p)^2}{npq}$$

and $$\gamma_1 = \sqrt{\beta_1} = \frac{(1-2p)}{\sqrt{npq}}$$

and $$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{npq\left[1+3(n-2)pq\right]}{n^2p^2q^2}$$

$$= \frac{1+3(n-2)\,pq}{npq}$$

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

and $$\gamma_2 = \beta_2 - 3 = \frac{1-6\,pq}{npq}$$

9.12.5 Recurrence Relation

We know that

$$P(X=r+1) = n_{c_r}\, p^{r+1}\, q^{n-(r+1)}$$

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

$$\therefore \quad \frac{P(X=r+1)}{P(X=r)} = \frac{n_{c_{r+1}}\, p^{r+1} q^{n-(r+1)}}{n_{c_r}\, p^r q^{n-r}}$$

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

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

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

Example 1. A Coin is tossed 5 times. What is the probability that head appears an odd number of times ?

Sol. Here probability of getting a head $P = 1/2$

Probability of not getting a head $q = 1 - 1/2 = 1/2$

Therefore the probability of getting odd numbers

$$= P(1) + P(3) + P(5)$$

$$= 5c_1\, pq^4 + 5c_3\, p^3 q^2 + 5c_5\, p^5$$

$$= 5\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^4 + \frac{5\times4\times3}{3\times2\times1}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^5$$

$$= \left(\frac{1}{2}\right)^5 [5+10+1]$$

$$= \frac{16}{32} = \frac{1}{2}$$

Example 2. If the sum of the mean and variance of a binomial distribution for 5 trials is 4.8. Find the distribution.

Sol. Since mean of the distribution $= np$

Variance of the binomial distribution $= npq$ and $n = 5$, therefore

$$np + npq = 4.8$$

or $$5p(1+q) = 4.8$$

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

$$50(1-q^2) = 48$$

or $$q^2 = 1 - \frac{48}{50} = \frac{2}{50}$$

$\Rightarrow$ $$q = \frac{1}{5}$$

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

Hence the required binomial distribution is

$$= (p+q)^n = \left(\frac{1}{5} + \frac{4}{5}\right)^n$$

Example 3. The incidence of occupational disease in an industry is such that the workers have a 20% chance of suffiereing from it. What is the probability that out of 6 workers 4 or more will suffer from disease ? **[Raj. B.E. EE-04]**

Sol. Probability of suffering worker from discease $P = 20\%$ i.e. $P = \frac{20}{100} = 1/5$

$\therefore$ Probability of not suffering worker from discease

$$q = 1 - \frac{1}{5} = \frac{4}{5}$$

Therefore the probability of 4 or more worker

$$P(X \geq 4) = P(4) + P(5) + P(6)$$

$$= {}^{6}C_{4}\ p^4 q^2 + {}^{6}C_{5}\ p^5 q + {}^{6}C_{6}\ p^6$$

$$= 15\left(\frac{4}{5}\right)^2\left(\frac{1}{5}\right)^4 + 6\left(\frac{4}{5}\right)\left(\frac{1}{5}\right)^5 + \left(\frac{1}{5}\right)^5 = \frac{53}{3125} \quad \textbf{Ans.}$$

Example 4. Calculate $P(r)$ for $r = 1,2,3,4$ and 5, taking $n = 5$ and $p = \frac{1}{6}$ with the help of the recurrence formula of the binomial distribution.

Sol. We know that the recurrence formula of Binomial distribution

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

Here $n = 5,\ p = \frac{1}{6},\quad q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}$

$$P(r+1) = \frac{5-r}{r+1} \cdot \frac{1}{5} P(r)$$

$\because$ $P(r) = {}^{n}C_{r}\ P^r q^{n-r}$

for $r = 0$, $P(1) = \frac{5}{1} \times \frac{1}{5} P(0) = q^5 = \left(\frac{5}{6}\right)^5 = \frac{3125}{7776}$

for $r = 1$, $P(2) = \frac{4}{2} \cdot \frac{1}{5} P(1) = 2 \times \frac{1}{5} \times 5 \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^4 = \frac{625}{7776} = 0.0803$

for $r = 2$, $P(3) = \frac{3}{3} \times \frac{1}{5}\ {}^{5}C_{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$

$$= \frac{1}{5} \times 5 \times 2 \times \frac{125}{7776} = \frac{250}{7776}$$

for $r = 3$, $P(4) = \frac{2}{4} \times \frac{1}{5} \cdot {}^{5}C_{3} \left(\frac{1}{6}\right)^{3} \left(\frac{5}{6}\right)^{2}$

$$= \frac{1}{2} \times \frac{1}{5} \times 10 \times \frac{25}{7776}$$

$$= \frac{25}{7776}$$

for $r = 4$, $P(5) = \frac{1}{5} \times \frac{1}{5} \times {}^{5}C_{4} \left(\frac{1}{6}\right)^{4} \left(\frac{5}{6}\right)$

$$= \frac{1}{25} \times 10 \times \frac{5}{7776}$$

$$P(5) = \frac{2}{7776}$$

and $P(6) = 0$

Example 5. The sum and product of the mean and variance of a binomial distribution are 24 and 128 respectively. Find the distribution.

Sol. According to given question

Mean + Variance = 24

Mean × Variance = 128

i.e. $np + npq = 24$ ... (1)

$np(npq) = 128$ ... (2)

$np(1+q) = 24$

$$np = \frac{24}{1+q}$$

$$\frac{24}{1+q}\left(\frac{24}{1+q}q\right) = 128 \quad \text{[from (2)]}$$

$$2q^2 - 59 + 2 = 0$$

$q = 2, \frac{1}{2}$

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

$\therefore$ (2) $\Rightarrow \quad n\left(\frac{1}{2}\right)\left(1+\frac{1}{2}\right) = 24$

$\Rightarrow \quad n = \frac{96}{3} = 32$

Hence the required Binomial distribution

$$= \left[\left(\frac{1}{2}+\frac{1}{2}\right)^{32} = (q+p)^n\right]$$

Example 6. Eight coins are tossed at 256 times. Find expected frequencies and also find the mean and variance.

Sol. Probability of success, $p = \frac{1}{2}$, probability of failure, $q = \frac{1}{2}$ and $n = 8,\ r = 256$.

The probability of getting r success

$$P(r) = n_{c_r}\, p^r q^{n-r}$$

$$= 8_{c_r}\left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{8-r}$$

$$= 8_{c_r}\left(\frac{1}{2}\right)^8 = 8_{c_r}\ \frac{1}{256}$$

Hence 8 coins are tossed 256 times, then frequency of r successes are

$$f(r) = 256 \times \frac{1}{256}\ {}^8C_r = {}^8C_r$$

i.e. 1, 8, 28, 56, 70, 56, 28, 8 and 1 for $r = 0, 1, 2, \ldots\ldots, 8$.

Mean of Binomial distribution $\mu = np$

$$= 8 \times \tfrac{1}{2} = 4$$

Variance $\quad \sigma^2 = npq$

$$= 8 \times \frac{1}{2} \times \frac{1}{2} = 2$$

Example 7. If m things are distributed among 'a' men and 'b' women, show that the probability that the number of things received by men is odd, is

$$\frac{1}{2}\left[\frac{(b+a)^m-(b-a)^m}{(b+a)^m}\right]$$

Sol. According to question

Number of things $= m$

Distributed number of men $= a$, number of women $= b$

Let P be the probability that a thing is received by man $p=\frac{a}{a+b}$ ... (1)

$$\Rightarrow \qquad q=1-p=1-\frac{a}{a+b}=\frac{b}{a+b} \qquad \text{... (2)}$$

The probability that out of m things exactly r are received by men and the remaining by women can be represented by binomial distribution.

$$P(X=r)=m_{c_r}\, p^r\, q^{m-r} \qquad r=0,1,2,\ldots\ldots m\frac{1}{2}$$

The probability R that the number of things received by men is odd is given by

$$R=P(1)+P(3)+P(5)+\ldots\ldots$$

$$={}^mc_1\, pq^{m-1}+{}^mc_3\, p^3q^{m-3}+{}^mc_5\, p^5q^{m-5}+\ldots.$$

but by the binomial theorem, we have

$$(q+p)^m-(q-p)^m=2\left[{}^mc_1\, q^{m-1}\, p+{}^mc_3\, q^{m-3}\, p^3+\ldots.\right]=2R$$

$$R=\frac{1}{2}\left[(q+p)^m-(q-p)^m\right]$$

Also $q+p=1$ and $q-p=\frac{b-a}{b+a}$

Therefore $$R=\frac{1}{2}\left[1-\frac{(b-a)^m}{(b+a)^m}\right]=\frac{1}{2}\left[\frac{(b+a)^m-(b-a)^m}{(b+a)^m}\right]$$

Example 8. Probability that a man aged 60 would be alive till the 70 years of age is 65%. Find the probability that at teast 7 out 10 such men would be alive till 70 years of age.

Sol. Given $n=10, \quad p=\frac{65}{100}=0.65, \quad q=1-p=0.35$

$\because \quad P(X=r)=n_{c_r}\, p^r q^{n-r}$

Required probability $P(X \geq 7)=P(7)+P(8)+P(9)+P(10)$

$$=10_{c_7}\, p^7 q^3+10_{c_8}\, p^8 q^2+10_{c_9}\, p^9 q+p^{10}$$

$$=10_{c_3}(0.65)^7(0.35)^3+10_{c_2}(0.65)^8(0.35)^2+10_{c_1}(0.65)^9(0.37)+(0.65)^{10}$$

$$=0.252+0.1755+0.0725+0.01346$$

$$=0.513$$

Example 9. If 10% of the pens manufactured by a company are defective, find the probability that a box of 12 pens contains.

(i) Exactly two defective pens.

(ii) At least two defective pens.

Sol. Probability of a defective pen $(p)=0.1$

$\therefore \quad q=1-p=0.9,$ and $n=12$

(i) Probability that the box contains exactly two defective pens.

$$=12_{c_2}\, p^2 q^{10}=12_{c_2}(0.1)^2(0.9)^{10}=0.2301 \quad \textbf{Ans.}$$

(ii) Probability that the box contains at least two defective pens = 1 – (Prob. that the box contain either 9 or 10 pens).

$$=1-\left[p(r=0)+p(r=1)\right]$$

$$=1-\left[12_{c_0}(0.9)^{12}+12_{c_1}(0.9)^{11}(0.1)\right]=0.341 \quad \textbf{Ans.}$$

Example 10 : Out of 800 familes with 4 children each, how many families would be expected to have

(a) 2 boys and 2 girls (b) at least one boy (c) at most 2 girls.

Sol. If a family having a boy is a success, then $p=\frac{1}{2}$ and for girl $(q)=\frac{1}{2}$.

Here number of families $(N) = 800$

(a) Probability for 2 boys and 2 girls is

$$p(x=2) = 4_{c_2}\ p^2q^2 = \frac{4.3.2.1}{2.1}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^4 = 3/8$$

$\therefore$ Number of families having 2 boys and 2 girls are equal to

$$N\left(\frac{3}{8}\right) = 800\times\frac{3}{8} = 300 \quad \textbf{Ans.}$$

(b) Probability for at least one boy is

$$= 1-p\,(0)$$

$$= 1-\left(\frac{1}{2}\right)^4 = \frac{15}{16}$$

$\therefore$ No. of families having at least one boy is equal to

$$N\left(\frac{15}{16}\right) = 800\left(\frac{15}{16}\right) = 750 \quad \textbf{Ans.}$$

(c) Probability for at most 2 girls is

$$= p(2)+p(3)+p(4)$$

$$= 1-\{p(0)+p(1)\}$$

$$= 1-\left[q^4+4pq^3\right]$$

$$= 1-\left[\left(\frac{1}{2}\right)^4+4\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3\right] = 1-\frac{5}{16} = \frac{11}{16}$$

$\therefore$ No. of families having at most two girls is equal to $N\left(\frac{11}{16}\right) = 800\left(\frac{11}{16}\right) = 550$ **Ans.**

9.13 POISSON DISTRIBUTION

It is a discrete probability distribution. It was developed by the french mathematician and Physicist **Simeon Denis poisson** in 1837, poisson distribution may be expected in case where the chance of any individual event being a success or failure is small. The distribution is used to describe the behavior of rare events such as the no. of persons born blind per year in a large city, no. of printing mistakes in a book, the no. of accidents on a road, etc. and has been called "the law of inprobable events".

The poisson distribution can be derived as a limiting case of the binomial distribution by making p very small and n very large and keeping average no. of occurrences of an event is fixed i.e. $np = m$.

The probability of r successes out of n trials in a binomial distribution is

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

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

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

Now as $n \to \infty$, $p \to 0$ and $np = m$ (fixed), we have

$$P(x=r) = \frac{m^{r}}{r!} \underset{n\to\infty}{at} \left[1-\frac{m}{n}\right]^{n-r}$$

$$P(X=r) = \frac{m^{r}}{r!} \underset{n\to\infty}{at} \frac{\left(1-\frac{m}{n}\right)^{n}}{\left(1-\frac{m}{n}\right)^{r}} = \frac{m^{r}}{r!}\ e^{-m}$$

$$\therefore \quad \boxed{P(X=r) = e^{-m}\ \frac{m^{r}}{r!}\ ;\quad r = 0, 1, 2, 3, \ldots.}$$

Which is the probability for the random variable $(X=r)$ in the poission distribution on substituting the values of r in poisson distribution formula, we get the probabilities of 0, 1, 2, ... r, ... sucesses in poisson distribution.

i.e.
$$\sum_{r=0}^{\infty} P(X=r) = e^{-m} + \frac{me^{-m}}{1!} + \frac{m^2}{2!}e^{-m} + \ldots + \frac{m^r}{r!}e^{-m} + \ldots$$

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

$$= e^{-m}\, e^{m}$$

$$= 1$$

Hence the total probability is one, which satisfies to be a probability distribution.

9.13.1 Mean and Variance of Poisson Distribution

The mean of the poisson distribution is given by

$$\mu = E(X) = \sum_{r=0}^{\infty} X_r\, P(X=r)$$

$$= \sum_{r=0}^{\infty} r\; P(X=r) = \sum_{r=0}^{\infty} r\; e^{-m}\, \frac{m^r}{r!}$$

$$= e^{-m} \sum \frac{e^{-m} m^2}{(r-1)!}$$

$$= e^{-m}\left[m + \frac{m^2}{11} + \frac{m^3}{2!} + \ldots\ldots\right]$$

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

$\therefore$ $$\mu = \boldsymbol{m e^{-m}\, e^{m} = m}$$

$\because$ Variance $$\left(\sigma^2\right) = E\left(X^2\right) - \left[E(X)\right]^2$$

$$= \sum_{r=0}^{\infty} r^2\; p(X=r) - m^2$$

$$= \sum_{r=0}^{\infty} \left[r(r-1) + r\right] p\,(X=r) - m^2$$

$$= \sum_{r=0}^{\infty} r\,(r-1)\,p(X=r) + \sum_{r=0}^{\infty} r\,p(X=r) - m^2$$

$$= \sum_{r=0}^{\infty} r\,(r-1)\,e^{-m}\,\frac{m^r}{(r-2)!} + m - m^2$$

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

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

$$= m^2\,e^{-m}\,e^{m} + m - m^2 = m^2 + m - m^2$$

Hence, $\boxed{Variance\,(\sigma^2) = m}$

$\Rightarrow$ $\boxed{Standard\ deviation\,(\sigma) = \sqrt{m}}$

9.13.2 Moment Generating Function of Poisson Distribution

The moment generating function about the origin for the poisson distribution is given by

$$M_X(t) = E\left[e^{tX}\right] = \sum_{r=0} e^{tX_r}\,p(X=r)$$

$$= \sum_{r=0}^{\infty} e^{tr}\,e^{-m}\,\frac{m^r}{r!}$$

$$= e^{-m}\left[\sum_{r=0}^{\infty} e^{tr}\,\frac{m^r}{r!}\right]$$

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

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

$$= e^{-m}\,e^{me^{t}} = e^{m\left(e^{t} - \ell\right)}$$

$$\Rightarrow \quad \boxed{M_X(t) = e^{(e^t-1)m}}$$

Now will obtain the moments about origin.

$$\mu_1' = \text{mean} \quad = \left[\frac{d}{dt} M_X(t)\right]_{t=0} = \left[\frac{d}{dt} e^{m(e^t-1)}\right]_{t=0}$$

$$= \left[me^t \; e^{m(e^t-1)}\right]_{t=0} = m$$

$$\Rightarrow \quad \boxed{\mu_1' = m}$$

Which is first moment about origin.

Now $\mu_2' = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0} = \left[\frac{d^2}{dt^2} e^{m(e^t-1)}\right]_{t=0}$

$$= m\left[\frac{d}{dt}\left\{e^t \; e^{m(e^t-1)}\right\}\right]_{t=0}$$

$$= m\left[e^t \; e^{m(e^t-1)} + me^{2t} \; e^{m(e^t-1)}\right]_{t=0}$$

$$= m\left[1+m\right] = m^2 + m$$

$$\therefore \quad \boxed{\mu_2' = m^2 + m}$$

Which represent the second moment about origin.

Again $\mu_3' = \left[\frac{d^3}{dt^3}\left(M_X|t|\right)\right]_{t=0}$

$$= m_4^1\left[\frac{d}{dt}\left\{e^t \; e^{m(e^t-1)} + me^{2t} \; e^{m(e^t-1)}\right\}\right]_{t=0}$$

$$\Rightarrow \mu_3' = m^3 + 3m^2 + m \text{ (on simplification)}$$

similarly $\mu_4' = m^4 + 6m^3 + 7m^2 + m$

9.13.3 Central Moments of Poisson Distribution

Moment about mean can be obtaine by using interrelation between moments are as follows:

$\mu_1 = 0$ for all distributions

$$\text{Variance} = \mu_2 = \mu_2' - \mu_1'^2 = m^2 + m - m^2 = m \quad (\because \mu_1' = m)$$

$$\mu_3 = \mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3$$

$$= (m^3 + 3m^2 + m) - 3m(m^2 + m) + 2m^3$$

$$= m^3 + 3m^2 + m - 3m^3 - 3m^2 + 2m^3$$

$$\mu_3 = m$$

and

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= (m^4 + 6m^3 + 7m^2 + m) - 4(m^3 + 3m^2 + m)m \quad +6(m^2 + m)m^2 - 3m^4$$

$$\mu_4 = 3m^2 + m$$

9.13.4 Karl Pearson Coefficients of Skewness and Kurtosis

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

$$\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}}$$

and

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

$$\gamma_2 = \beta_2 - 3 = \frac{1}{m}$$

9.13.5 Recurrence Formula for Poisson Distribution

$\because$

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

and

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

$\therefore$

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

$\Rightarrow$

$$P(X = r+1) = \frac{m}{r+1}P(X = r)$$

Example 11. Suppose on an average 1 house in 1000 in a certain distinct has a fire during a year. If there are 2000 houses in that district, what is the probability that exactly 5 houses will have a fire during the year ? (Given $e^{-2} = 0.1353$) **[Raj. B.E. (CS) 2003, 2006]**

Sol. Here $n = 2000,\ P = \dfrac{1}{1000}$

$$\therefore \quad m = np = \frac{2000}{1000} = 2$$

Also we have

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

therefore $$P(5) = \frac{e^{-2} 2^5}{\lfloor 5} = (0.1353)\frac{2^5}{\lfloor 5}$$

$$= \frac{0.1353 \times 32}{120}$$

$$P(5) = 0.036$$

Example 12. A manufacturer of cotter pins known that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins. Will be defective. What is the approxmate probability that a box will fails to meet the guaranteed quality.

[Raj. B.E. (CS) 2006]

Sol. Given $n = 100$

probability of defective pin $= 5\% \dfrac{5}{100} = 0.05$

Therefore mean of the defective pins $= np$

$$m = 100 \times .05 = 5$$

so poisson distribution

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

$$= \frac{e^{-5} 5^r}{\lfloor r}$$

Probability that a box will fail to meet the guaranteed quality

$$P(X \geq 10) = 1 - P(x \leq 10)$$

$$= 1 - e^{-5} \sum_{r=0}^{10} \frac{5^r}{\lfloor r}$$

Example 13. If X has a poisson distribution such that $P(X=1) = 1 = P(X=2)$, find $P(X=4)$

Sol. Here we know that poisson distribution

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

$$P(X=1) = \frac{e^{-m}m^1}{\lfloor 1} = me^{-m}$$

$$P(X=2) = \frac{e^{-m}m^2}{\lfloor 2}$$

since $P(x=1) = P(X=2)$

$$me^{-m} = \frac{e^{-m}m^2}{2} \Rightarrow m=2$$

therefore $P(X=4) = \frac{e^{-2}2^4}{\lfloor 4}$

$$= \frac{e^{-2}\times 16}{4\times 3\times 2\times 1} = \frac{2}{3e^2}$$

$$P(x=4) = \frac{2}{3e^2}$$ **Ans.**

Example 14. Razor blades are supplied by a manufacturing company in packet of 10. There is a probability of 1 in 100 blades to be defective. Using poisson distribution calculate the number of packets containing one defective blade, no defective blade and all defective blades in a consignment of 10,000 packets ?

Sol. Here $P = 0.01$, $n = 10$, and $m = np = 0.1$

using poisson distribution formula

$$P(x=r) = e^{-m}\frac{m^r}{r!}, r = 0,1,2,\ldots\infty$$

(a) Probability that a packet of 10 blades contains one defective blade is

$$P(r=1) = e^{-0.1}(0.1) = 0.09$$

$\therefore$ No. of packets containing one defective blade out of 10,000 blades manufactured is

$$= 10,000\times .09 = 900$$

(b) Prob. that a packet contains no defective blade is

$$P(r=0) = e^{-0.1} = 0.90483$$

$\therefore$ No. of packets containing no defective blade is equal to

$$10,000 \times 0.90483 = 9048 \quad \text{Ans.}$$

(c) The prob that a packet contains all defective blades

$$P(r=10) = e^{-m}\frac{m^{10}}{10!} = \frac{e^{0.1} \times (0.1)^{10}}{10!}$$

$$= 2.4934 \times 10^{-17}$$

Hence no. of packets containing all defective blades is equal to

$$10,000 \times 2.4934 \times 10^{-17} = 2.4934 \times 10^{-13} = 0 \quad \text{Ans.}$$

Example 15. A skilled typist, on routine work, kept a record of mistakes made per day during 300 working days as :

Mistakes per days	0	1	2	3	4	5	6
No of days	143	90	42	12	9	3	1

Compute the frequencies of the poisson distribution which has the same mean and total frequency as the above distribution?

Sol. $\because$ We know that the mean of poisson distribution $(m) = \frac{\sum fx}{\sum f} = \frac{267}{300} = 0.89$

therefore theretical frequencies $= 300P(x=r)$

$= 300e^{-m}\frac{m^r}{r!}$ is as follows.

x	$P(x)$	$300 \times P(x)$
0	0.411	123
1	0.365	110
2	0.163	49
3	0.048	14
4	0.011	3
5	0.02	1
6	0.003	0

CONTINUOUS PROBABILITY DISTRIBUTION

9.14 INTRODUCTION

So far we have dealt with discrete distributions like binomial and poisson distribution, they are relate to the occurance of distinct events. In order to have mathematical distribution suitable for dealing with quantities whose magnitude is continuously variable, like temperature, heights, weights etc a continuous distribution is needed. The normal distribution happens to be most useful theoretical distribution for continuous variable. The normal distribution was first discovered in 1733 by english mathematician **De-Moiver**, who obtained this continuous distribution as a limiting case of the binomial distribution. The probability of a continuous random variate X, which is defined in the given domain (a, b) can be find by using the probability density function, which is as follows :

Suppose $f(x)$ be a continuous function,

then $E[x] = \int_{-\infty}^{\infty} xf(x)dx = \bar{x}$

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

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

$f(x)$ **is called probability density function if :**

(i) $f(x) \geq 0 \quad \forall x \in [a,b]$

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

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

9.14.1 Normal Distribution or Gaussian Distribution

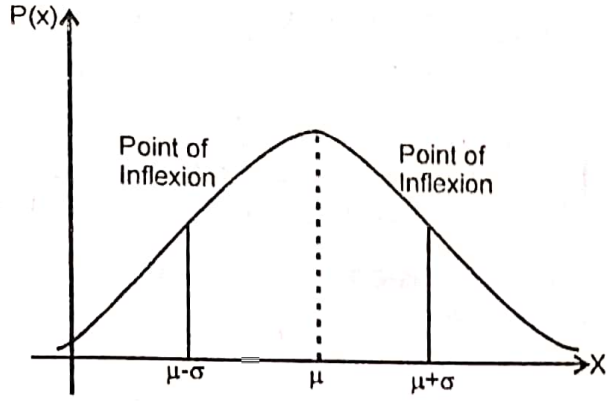
Normal distribution is continuous probability distribution. It is derived as the limiting form of the Binomial distribution for large value of n i.e. $n \to \infty$ and p and q are not very small.

Definition : A continuous random variable X is said to have a normal distribution with parametic μ (mean) and σ^2 (variance). If its probability density function is given by :

$$f(x) = f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} ; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

where X is a normal variate as is expressed as $N(\mu;\sigma)$

The graph of the normal frequency function is shown infigure (1). This graph is called the normal probability curve. It is symmetrical about the ordinate $x=\mu$

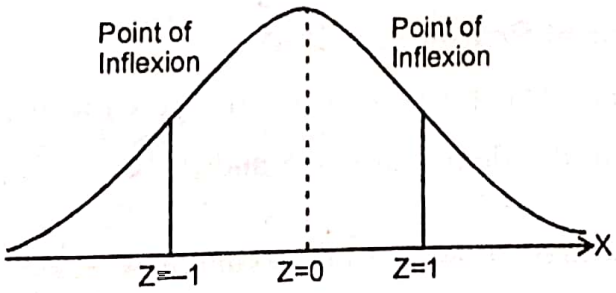


(Fig. 1)

If x is normal random variable with mean μ and standard deviation σ, then the random variable $Z=\frac{x-\mu}{\sigma}$ is called the **standard normal variable.** If mean $\mu=0$ and standard deviation $\sigma=1$ then normal distribution

$$P(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}},\ Z \text{ being any real}$$

The graph of the standard form of the normal frequncy function is symmetrical about $Z=0$ and It can be proved by methods of calculation that it has two paints of inflexion, one on either side of $Z=0$ and these point of inflexion correspond to $Z=-1$ and $Z=1$. The graph of standard form of normal frequency function is showin figures (2)



(Fig. 2)

Normal distributed is also denoted by $N(\mu,\sigma)$ Now we will prove that the normal distribution function $f(x)$ is a probability density function.

(i) Let $I=\int_{-\infty}^{\infty} f(x)dx$

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

Let $$Z = \frac{x-\mu}{2} \Rightarrow dz = \frac{dx}{\sigma}$$

then $$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{z^2}{z}} dz$$

Put $$\frac{z^2}{2} = t \Rightarrow zdz = dt$$

Now $$I = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-t} . \frac{dt}{\sqrt{2t}} = \frac{1}{\pi} \int_{0}^{\infty} e^{-t} t^{-1/2} dt$$

$$= \frac{1}{\sqrt{\pi}} \overline{\left|\left(\frac{1}{2}\right)\right.} \qquad \left[\because \overline{|(n)} = \int_{0}^{\infty} e^{-t} t^{n-1} dt\right]$$

$$= \frac{1}{\sqrt{\pi}}^{\sqrt{\pi}} \qquad \left[\because \overline{\left|\left(\frac{1}{2}\right)\right.} = \sqrt{\pi}\right]$$

$$\therefore \qquad \int_{-\infty}^{\infty} f(x)dx = 1$$

Hence the given function $f(x)$ for $N(\mu,\sigma)$ satisfies both the conditions, then it is said to be. Probability density functions.

9.14.2 Properties of Normal Probability Curve

[Raj. BE CS (Back) 07 (Old)]

(i) The curve is symmetrical about the mean, therefore mean and median coincide.

(ii) There are two point of inflexion at $\mu+\sigma$ and $\mu-\sigma$.

(iii) The height of a normal curve is at its maximum $= \frac{1}{\sigma\sqrt{2\pi}}$ at the mean then mean and mode concide.

(iv) Total area under the normal curve is unity.

(v) Normal curve is unimodal i.e. it has only one mode.

(vi) Curve is asymptotic to X-axis.

(vii) Mean, mode and medion of the distribution coincide i.e. mean = mode = median.

(viii) The Area under the normal curve is distributed as follows :

(a) Between $\mu-\sigma$ and $\mu+\sigma$, it is 68.27%

(b) Between $\mu-2\sigma$ and $\mu+2\sigma$ it is 95.45%

(c) Between $\mu-3\sigma$ and $\mu+3\sigma$ it is 99.73%.

Probability density function (PDF) of the standard normal variate Z then

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-Z^2/2}$$

(ix) The shape characterisation coefficients for this distribution are $\beta_1=0$, $\beta_2=3$ i.e. $\beta_1=0$ shows the symmetry.

9.14.3 Mean of Normal Distribution

Mean $\mu_1' = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, du$

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

Let $Z=\dfrac{x-\mu}{\sigma} \Rightarrow x=\mu+\sigma z \Rightarrow dx=\sigma dz$

Therefore, Mean

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\mu+\sigma z)\, e^{-z^2/2}\, dz$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mu e^{-z^2/2}dz+\sigma\int_{-\infty}^{\infty} Z\, e^{-z^2/2}\, dz$$

$$=\frac{2\mu}{\sqrt{2\pi}}\int_{0}^{\infty} e^{-z^2/2}dz+0$$

$$\left[\because Ze^{z^2/2} \text{ and } e^{-z^2/2} \text{ is odd function is even function}\right]$$

$$=\frac{2\mu}{\sqrt{2\pi}}\int_{0}^{\infty} e^{-z^2/2}\, dz$$

$$= Put\ \frac{z^2}{2}=t \Rightarrow zdz=dt$$

$$= \mu\sqrt{\frac{2}{\pi}} \int_0^\infty e^{-t} \frac{dt}{\sqrt{2t}}$$

$$= \frac{\mu}{\sqrt{\pi}} \int_0^\infty e^{-t} t^{\frac{1}{2}-1} dt$$

$$= \frac{\mu}{\sqrt{\pi}} \left\lfloor \frac{1}{2} = \frac{\mu}{\sqrt{\pi}} \sqrt{\pi} = \mu \right.$$

$$\boxed{\textit{Mean} = \mu}$$

Mean of the normal distribution is μ

9.14.4 Variance of Normal Distribution

Variance $\sigma^2 = \mu_2 = \int_{-\infty}^{\infty} E\left(X - \bar{X}\right)^2 = E\left(X - \mu\right)^2$

$$= \int_{-\infty}^{\infty} E\left(x - \mu\right)^2 f\left(x\right) dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(x - \mu\right)^2 e^{-\left(\frac{x-\mu}{2\sigma^2}\right)^2} dx$$

Let $\frac{x - \mu}{\sigma} = Z$

$\Rightarrow$ $dx = \sigma dz$

Variance $\left(\mu_2\right) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\sigma z\right)^2 e^{-z^2/2} \sigma\, dz$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_0^{\infty} z^2 e^{-z^2/2} dz$$

Let $\frac{Z^2}{2} = t \Rightarrow zdz = dt$

$$\therefore \quad dz = \frac{dt}{\sqrt{2t}}$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_0^\infty 2te^{-t} \frac{dt}{\sqrt{2t}}$$

$$\frac{2\sigma^2}{\sqrt{\pi}} \int_0^\infty 2t \; e^{-t} t^{1/2} dt$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_0^\infty e^{-t} \; t^{3/2-1} \; dt$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \left|\overline{\left(\frac{3}{2}\right)}\right.$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \; \frac{1}{2} \left|\overline{\left(\frac{1}{2}\right)}\right.$$

$$= \frac{\sigma^2}{\sqrt{\pi}} \times \sqrt{\pi} = \sigma^2$$

$\therefore$ Variance σ^2 and standard deviation σ

9.14.5 Moment Generating Function of Normal Distribution

The moment generating function of normal variate X about origin is given by

$$M_X(t) = E\left(e^{tx}\right) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} \; e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Let $$= \frac{x-\mu}{\sigma} = z \Rightarrow dx = \sigma\, dz$$

$\Rightarrow$ $$M_X(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(\mu+\sigma z)} \; e^{-\frac{z^2}{2}} \; \sigma dz$$

$$= \frac{1}{\sqrt{2\pi}} e^{\mu t} \int_{-\infty}^{\infty} e^{\sigma z t - \frac{z^2}{2}} dz$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\sigma z t - \frac{z^2}{2}} dz$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left\{z^2 - 2\sigma z t + \sigma^2 + t^2\sigma^2 - \sigma^2 t^2\right\}} dz$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left\{(z-\sigma t)^2 - \sigma^2 t^2\right\}} dz$$

$$= \frac{e^{\mu t + \frac{\sigma^2 t^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv \qquad \left[\begin{array}{l} Let\ z - \sigma t = v \\ \therefore\ dz = dv \end{array}\right]$$

$$= \frac{e^{\mu t + \frac{\sigma^2 t^2}{2}}}{\sqrt{2\pi}} \cdot 2 \int_{0}^{\infty} e^{-\frac{v^2}{2}} dv$$

$$= \frac{e^{\mu t + \frac{\sigma^2 t^2}{2}}}{\sqrt{2\pi}} \cdot 2 \sqrt{\frac{\pi}{2}}$$

$$\boxed{M_X = e^{\mu t + \frac{\sigma^2 t^2}{2}}}$$

Which is the moment generating function of normal variate.

Now we will find moments about origin from $M_X(t)$

$\because$ $\mu_r = e^{\mu t + \frac{\sigma^2 t^2}{2}}$

$$\mu_1' = \left[\frac{d}{dt} M_X(t)\right]_{t=0}$$

$$= \frac{d}{dt}\left[e^{\mu t + \frac{\sigma^2 t^2}{2}}\right]_{t=0} = \left[\left(\mu t + \sigma^2 t\right) e^{\mu t + \frac{\sigma^2 t^2}{2}}\right]_{t=0}$$

$$\mu_1' = \mu + 0\ e^0 = \mu$$

$$\therefore \quad \textbf{Mean} = \mu_1' = \mu$$

$$\mu_2' = \left[\frac{d^2}{dt^2}\left(e^{\mu t + \frac{\sigma^2 t^2}{2}}\right)\right]_{t=0}$$

$$= \left\{\frac{d}{dt}\left[\left(\mu + \sigma^2 t\right) e^{\mu t + \frac{\sigma^2 t^2}{2}}\right]\right\}_{t=0}$$

$$= \left[\sigma^2 e^{\mu t + \frac{\sigma^2 t^2}{2}} + \left(\mu + \sigma^2 t\right)^2 e^{\mu t + \frac{\sigma^2 + t^2}{2}}\right]_{t=0}$$

$$\mu_2' = \sigma^2 e^0 + \mu^2 e^0 = \sigma^2 + \mu^2$$

and variance $\mu_2' = \sigma^2 + \mu^2$

$$\mu_3' = \left[\frac{d^3}{dt^3}\left(e^{\mu t + \frac{\sigma^2 t^2}{2}}\right)\right]_{t=0}$$

$$= \left[\frac{d}{dt}\left(\sigma^2 e^{\mu t + \frac{\sigma^2 t^2}{2}} + \left(\mu + \sigma^2 t\right)^2 e^{\mu t + \frac{\sigma^2 t^2}{2}}\right)\right]_{t=0}$$

$$= \left[\sigma^2\left\{\left(\mu + \sigma^2 t\right) e^{\mu t + \frac{\mu^2 t^2}{2}}\right\} + 2\sigma^2\left(\mu + \sigma^2 t\right) e^{\mu t + \frac{\sigma^2 t^2}{2}} + \left(\mu + \sigma^2 t\right)^3 e^{\mu + \frac{\sigma^2 t^2}{2}}\right]_{t=0}$$

$$= \sigma^2\mu + 2\sigma^2\mu + \mu^3$$

$$\mu_3' = 3\sigma^2\mu + \mu^3$$

$$\mu_4' = 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4$$

Alternate Method :

Moment generating function $M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}} = e^{t\left(\mu + \frac{\sigma^2 t}{2}\right)}$

$$M_X(t) = 1 + \frac{t}{\lfloor 1}\left(\mu + \frac{\sigma^2 t}{2}\right) + \frac{t^2}{\lfloor 2}\left(\mu + \frac{\sigma^2 t}{2}\right)^2 + \frac{t^3}{\lfloor 3}\left(\mu + \frac{\sigma^2 t}{2}\right)^3 + \frac{t^4}{\lfloor 4}\left(\mu + \frac{\sigma^2 t}{2}\right)^4 + \ldots..$$

$\therefore$ $E(X) = mean = coefficient\ of\ \frac{t}{\lfloor 1} = \mu$

$E(X^3) = variance = coefficient\ of\ \frac{t^2}{\lfloor 2} = \sigma^2 + \mu^2$

$E(X^3) = coefficient\ of\ \frac{t^3}{\lfloor 3} = 3\mu\sigma^2 + \mu^3$

9.14.6 Moment Generating Function about Mean of Normal Variate X

Moment generating function about mean :

$$M_{\bar{X}}(t) = E\left(e^{t(X-\bar{X})}\right) = E\left(e^{t(X-\mu)}\right)$$

$$= e^{-\mu t}\ E\left(e^{tx}\right)$$

$$= e^{-\mu t}\ M_X(t)$$

$$= e^{-\mu t}\ e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

$$\boxed{M_{\bar{X}}(t) = e^{\frac{\sigma^2 t^2}{2}}}$$

Central moments of normal variate x can be find as

$$\mu_1 = \left[\frac{d}{dt} M_X(t)\ about\ mean\right]_{t=0} = \left[\frac{d}{dt} M_{\bar{X}}(t)\right]_{t=0}$$

$$= \left[\frac{d}{dt}\left(e^{\frac{\sigma^2 t^2}{2}}\right)\right]_{t=0}$$

$$= \left[\sigma^2 t \; e^{\frac{\sigma^2 t^2}{2}} \right]_{t=0} = 0$$

$$\mu_2 = \frac{d^2}{dt^2} M_{\bar{X}}(t) = \left\{ \frac{d}{dt} \left[\sigma^2 t \; e^{\frac{\sigma^2 t^2}{2}} \right] \right\}_{t=0}$$

$$\therefore \quad \mu_2 = \left\{ \sigma^2 \left[e^{\frac{\sigma^2 t^2}{2}} + \sigma^2 t^2 \; e^{\frac{\sigma^2 t^2}{2}} \right] \right\}_{t=0} = \sigma^2$$

$$\mu_3 = 0$$

and $$\mu_4 = 3\sigma^4$$

9.14.7 Karl Pearson's Coefficients β and γ for Normal Distribution

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

$$\beta_1 = 0$$

Since $$\gamma_1 = \sqrt{\beta_1} = 0$$

Thefefore curve is symmetric

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3$$

$$\beta_2 = 3$$

Hence curve is a normal curve

9.14.8 Fitting of Normal Distribution

We consider to fit a normal distribution to a given frequency distribution x_i and f_i $i = 1, 2, \ldots . n$ then we find mean and variance by

$$\mu = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \quad \text{and} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - \frac{\left(\sum_{i=1}^{n} f_i x_i \right)^2}{\sum_{i=1}^{n} f_i}}$$

From the given data

Hence the normal curve fitted to the given data is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

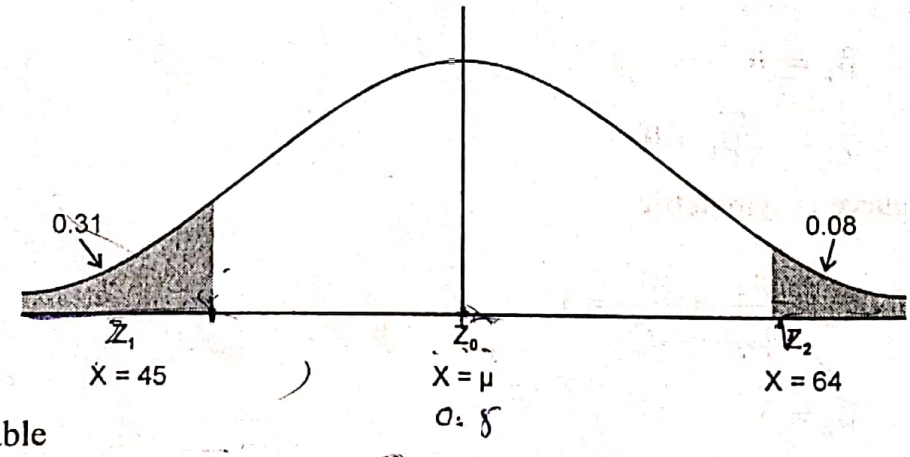
Example 16 : In a normal distribution 31% items are under 45 and 8% are above 64. Find the mean and standard deviation of the distribution.

Sol. Suppose μ and σ be the mean and S.D. respectively.

Given $P(X \le 45) = 0.31$ and $P(X > 64) = 0.08$

$$0.5 - \phi\left(\frac{X-\mu}{\sigma}\right) = 0.31 \quad ; \quad 0.5 - \phi\left(\frac{X-\mu}{\sigma}\right) = 0.08$$

$$\Rightarrow \quad \phi\left(\frac{45-\mu}{\sigma}\right) = 0.19 \quad ; \quad \phi\left(\frac{64-\mu}{\sigma}\right) = 0.42$$



from the table

$$P(0 < Z < 1.41) = 0.42 \quad ; \quad P(0 < Z < 0.5) = 0.19$$

$$\therefore \quad \frac{45-\mu}{\sigma} = -0.5 \quad \text{(Lies on negative side)}$$

and $$\frac{64-\mu}{\sigma} = 1.41 \quad \text{(Lies on negative side)}$$

on solving the above equations for μ and σ, we get

$\mu = 50$ and $\sigma = 10$ (approx) **Ans.**

Example 17 : The distribution of weekly weges for 500 workers in a factory is approximately normal with the mean and S.D. of Rs. 75 and Rs. 15 respectively. Find the number of workers who receive weekly wages.

(i) More than Rs. 90 (ii) Less than Rs. 45

Sol. Given $N = 500, \mu = 75$ and $\sigma = 15$

Let $$Z = \frac{x-\mu}{\sigma} = \frac{X-75}{15}$$

(i) $P(x > 90) = P(Z \geq 1) = 0.5 - P(0 < Z < 1)$

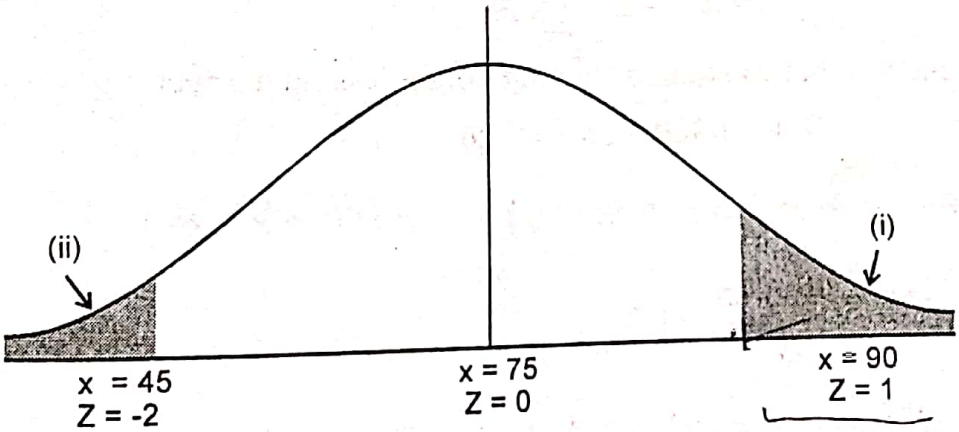
$= 0.5 - 0.3413$

$= -0.1587$

Hence number of workers $= 500 \times 0.1587$

$= 79.35$

$\simeq 79$ **Ans.**



(ii) For $P(x < 45) = P(Z < -2) = 0.5 - P(-2 \leq Z < 0)$

$= 0.5 - P(0 < Z \leq 2)$

$= 0.5 - 0.4772$

$= 0.0228$

$\therefore$ No. of workers $= 500 \times 0.0228 = 11.4$

$\simeq 11$ **Ans.**

Example 18. Define the normal distribution if the height of 300 students are normally distributed with mean 66.5 inches and standard deviation 3.3 inches. How many students have height (i) less than 5 feet (ii) between 5 feet and 5 feet 9 inches. Also find the height between which 99% of the student lie **[Raj. CP. 2004]**

Sol. We know that the normal distribution

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\quad,\ -\infty<x<\infty$$

Here x denote the height of the students.

Number of student N = 300

Here $\mu=66.5$ *and* $\sigma=3.3$

(i) $x=5$ *feet* $=5\times12=60$ *inches* , $Z=\dfrac{x-\mu}{\sigma}=\dfrac{60-66.5}{3.3}=-1.36$

$$P(X\le5\ \textit{feet})=P(X<60\ \textit{inches})=P(Z<-1.36)$$

$$=0.5-P(-1.36<Z<0)$$

$$=0.5-P(0<Z<1.36)$$

$$=.5-0.4131=0.0869$$

Hence the number of students having height less then 5 feet

$$=300\times.0869=26.70=26$$

(ii) $P(5\ \textit{feet}<X<5\ \textit{feet}\ 9\ \textit{inches})=P(60<x<69)$

and $Z=\dfrac{69-66.5}{3}=1.36$

$$P(60<x<69)=P(-1.36<Z<1.36)$$

$$=P(-1.36<Z<0)+P(0<Z<1.36)$$

$$=P(0<Z<1.36)+P(0<Z<1.36)$$

$$=2P(0<Z<1.36)$$

$$=2\times.4131=0.8262$$

Hence the number of students having height between 5 feet and 5 feet 9 inches = $300\times.8262=2485$

(iii) We have to find the height between which 99% of the student lie. Let students heights lie between x_1 and x_2. The curve is symmetric about $z=0$, let the height lies between $Z=\pm Z_1$

$$P(-Z_1 < Z < Z_1) = 99\% = 0.99$$

$$P(0 \le Z < Z_1) + P(0 < Z < Z_1) = .99$$

$$2P(0 < Z < Z_1) = .99$$

$$P(0 < Z < Z_1) = .495$$

We see the normal table if $Z_1 = 2.57$

$$-Z_1 = \frac{x_1 - \mu}{6} = -2.57 = \frac{x_1 - 66.5}{3.3}$$

$$x_1 = 66.5 - 8.481$$

$$x_1 = 58.019 \cong 58$$

and also

$$Z_1 = \frac{x_2 - \mu}{\sigma} = 2.57 = \frac{x_2 - 66.5}{3.3}$$

$$x_2 = 66.5 + 8.481$$

$$x_2 = 74.981$$

$$x_2 \cong 75$$

Hence height of 99% of the students lie between 58 inches and 75 inches.

Example 19. The mean length of steel bars produced by a company is 10 meters with standard deviation 20 cm. 5000 bars are purchased by a contractor. How many of these bars are expected to be shorter than 9.75 meters in length ? Assuming that the length of steel bars are normally distributed, use the following extract from the table of Areas from mean to a distante of $\frac{x - \bar{x}}{\sigma}$ from the mean under the curve

$\frac{x - \bar{x}}{\sigma}$	1.10	1.15	1.20	1.25	1.30
Area	0.3643	0.3749	0.3849	0.3944	0.4032

[Raj. BE (ME), 05]

Sol. According to question $\bar{x} = \mu = 10$

$$\sigma = 0.2, x = 9.75 \text{ and } N = 5000$$

therefore $$Z = \frac{x-\mu}{\sigma} = \frac{9.75-10}{0.20} = -1.25$$

We see the given table the area of the left of $Z = -1.25$ is $= 0.3944$

Therefore the Area from

$$Z = 0 \text{ to } Z = -1.25$$
$$= 0.5 - 0.3944$$
$$= 0.1056$$

Thus the ratio of steel bars of length less the $9.75m = 0.1056$

$\therefore$ Their number = Ratio $\times \mu$

$$= 0.1056 \times 5000 = 528$$

Example 20. For a certains normal distribution the first moment about 10 is 40 and that the 4th moment about 50 is 48, what is A.M. and S.D. of the distribution ?

Sol. Let μ and σ be A.M. and S.D.

then $\mu_1'(10) = 40 \Rightarrow E[x-10] = 40$

or $E(x) = 50 \ \therefore \mu = 50$ **Ans.**

Also $\mu_4 = 48 \Rightarrow 3\sigma^4 = 48 \ [\because \beta_2 = 3]$

Hence $\sigma = 2$ **Ans.**

Example 21. A coin is tossed 12 times. Find the probability both exectly and by fitting a normal distribution of getting (i) 4 heads (ii) at most 4 heads.

Sol. Suppose X be the randam variable denotes the no. of head obtained, than

$$P(H) = \frac{1}{2} \text{ and } P(\bar{H}) = P(T) = q = \frac{1}{2}, n = 12$$

for exact probability, using Binomial distribution $P(X = r) = n_{c_r} p^r q^{n-r}$

(i) $$P(X = 4) = {}^{12}C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^8 = 495 \left(\frac{1}{2}\right)^{12} = 0.1208$$

(ii) $P(X \le 4) = P(0)+P(1)+P(2)+P(3)+P(4)$

$$= \left(\frac{1}{2}\right)^{12}\left[1+{}^{12}C_1+{}^{12}C_2+{}^{12}C_3+{}^{12}C_4\right]$$

$$= \left(\frac{1}{2}\right)^{12}[1+12+66+220+495]$$

$$= 0.1937$$

Now we find the probability by fitting of normal distribution

$\because$ $\mu = np = 12\times\frac{1}{2} = 6$

and S.D. $\sigma = \sqrt{npq} = \sqrt{12\times\frac{1}{2}\times\frac{1}{2}} = \sqrt{3}$

(i) Prob. of 4 heads $= P(3.5 < X < 4.5)$

for $x = 3.5, z = \frac{3.5-6}{\sqrt{3}} = \frac{-2.5}{\sqrt{3}} = -1.44$

and for $x = 4.5, Z = \frac{4.5-6}{\sqrt{3}} = \frac{-1.5}{\sqrt{3}} = -0.866$

$\therefore$ $P(3.5 < X < 4.5) = P(-1.44 < Z < -0.866)$

$= P(-1.44 < Z < 0) - P(0 < Z < 0.866)$

$= P(0 < Z < 1.44) - P(0 < Z < 0.866)$

$= 0.4251 - 0.3051$

$= 0.12$

(ii) Probability for at most 4 heads

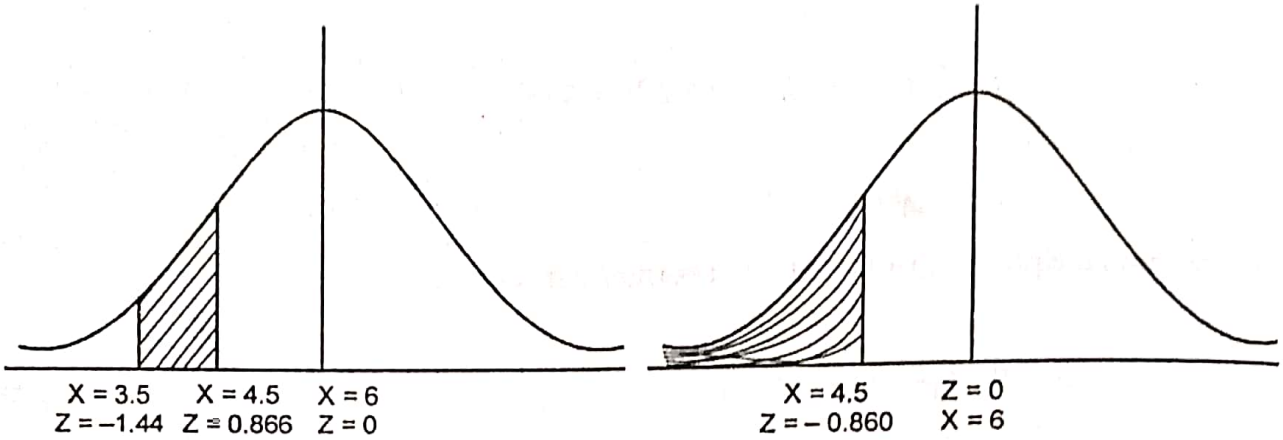
$= P(X < 4.5)$

$= P(Z < -0.866)$

$$= 0.5 - P(0 < Z < 0.866)$$

$$= 0.5 - 0.3051$$

$$= 0.1949$$



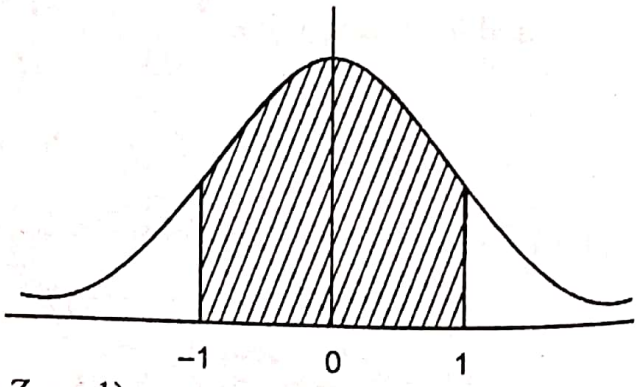
Example 22.A manufacturer knows from experience that the resistance of resistors he produces is normal with mean μ = 100 ohms and standard deviation $\sigma = 2$ ohms. What precentage of resistors will have resistance between 98 ohms and 102 ohms ?

Sol. Hence $\mu = 100$ ohms, $\sigma = 2$ ohms, $x_1 = 98$ and $x_2 = 102$

$$\because \quad Z = \frac{x - \mu}{\sigma}$$

$$\therefore \quad Z_1 = \frac{x_1 - \mu}{\sigma} = \frac{98 - 100}{2} = -1$$

and $$Z_2 = \frac{x_2 - \mu}{\sigma} = \frac{102 - 100}{2} = 1$$



Area between $Z_1 = -1$ and $Z_2 = 1$

= (Area between $Z = 0$ and $Z = -1$)

+ (Area between $Z = 0$ and $Z = -1$)

= 2 (Area between $Z = 0$ and $Z = 1$)

$= 2 \times 0.3413 = 0.6826$

Hence, percentage of resister having resistance between 98 ohms and 102 ohms is **68.26.**

EXERCISE – 9.2

Q.1 A coin is tossed 4 times. What is the probability of getting (a) two heads (b) Atleast two heads.

Ans. **(a) 0.375, (b) 0.687**

Q.2 Six dice are thrown 729 times. How many times do you expect atleast three dice to shown a five or a six ?

Ans. **233**

Q.3 An unbiased dice is thrown again and again until three sixes are obtained. Find the probability of obtaining the third six in the six throws of the dice.

Ans. **1250/46656**

Q.4 If 10% of the bolts produced by a machine are defective, use binomial distribution to determine the probability that out of 5 bolts selected at random, atmost one will be defective.

Ans. **0.918**

Q.5 Find the total number of trials when the mean number of trials is 9 and variance is 6.

Ans. $n = 27$

Q.6 The reaction of a special type of injection to a person has probability 0.001. Find the probability of reaction out of 2000 persons on : $[e^{-2} = 0.1353]$ **[Raj. BE II, 06]**

(a) 3 persons (b) more than 2 persons.

Ans. **0.1804, 0.3235**

Q.7 A pair of dice is thrown 200 times. If getting a sum of 9 is considered a success, find the mean and the variance of the number of successes.

Ans. **Mean = Variance = 22.22**

Q.8 A book contains 100 misprints distributed randamly through out its 100 pages. What is the probability that a page observed at random contains at least two misprints (Assume poisson Distribution)

Ans. **0.264**

Q.9 Using poisson's distribution, find the probability that the aces of spades will be drawn from a pack of well shaffted cards atleast once in 104 consecutive trials $(e^{-2} = 0.1353)$

Ans. **0.8647**

Q.10. A and B shoot independently until each has hit his own target. The probabilities of their hitting the target at cach shot are 3/5 and 5/7 respectively. Find the probability that B will require more shots than A.

Ans. **6/31**

Q. 11. Fit a poisson distribution for the following distribution

x	0	1	2	3	4	5
f	142	156	69	27	5	1

Q.12 A manufacturer knows that the condensers he makes contain on the average 1% of defectives. He packs then in boxes of 100. What is the probability that a box, picked at random will contain 3 or more faulty condeners ? $[e^{-1} = 0.3679]$

Ans. $P(r \geq 3) = 0.08025$

Q.13 If x is a normal variable with mean 11 and standard deviation 1.5, find the number 'a' such that $P(X > a) = 0.09$.

Ans. $x = 13.01$

Q.14 The marks obtained by a group of students who appeared for a test were normally distributed with mean 80 and standard deviation 6. Find the standard scores for the students who scored (i) 90 marks, (ii) 68 marks, (iii) 57 marks and (iv) 120 marks.

Ans. (i) 1.67, (ii) –2, (iii) –3.83, (iv) =3.67

Q.15 Assume the mean height of students in an exactly normal distribution to be 68.22 inches with a variance of 10.8 inches. How many students in a college of 1000 students would you expect to be over six feet tall ? (Area under a normal curve corresponding to Z = 1.15 is 0.3749).

Ans. 125

Q.16 For a certain normal distribution the firt moment about 10 is 40 and that the 4th moment about 50 is 48, what is the arithmetic mean and of the ditribution ?

$$[\mu_1' = 40, \beta_2 = 3, \beta_2 = \frac{\mu_4}{\sigma^4} \Rightarrow \mu_4 = 3\sigma^4, \sigma = 2]$$

Q.17 If X and Y are independent random variables having normal distributions with a common mean μ with variances 4 and 48 respectively, such that $P(X + 2Y < 3) = P(2X - Y \geq 4)$ Determine μ.

Ans. 2.1

Q.18 A sample of 100 dry battery cells tested to find the length of life produced the following result $\mu = 12$ Hrs. and $\sigma = 3$ Hrs. Assuming the data to be normally distributed, what percentage of batery cells are expected to have life :

(a) More than 15 Hrs. (b) Between 10 and 14 Hrs.

(c) Less than 6 Hrs.

Ans. : (a) 15.87%, (b) 49.74%, (c) 2.28%

□□□

1. The uniform distribution

The Uniform or Rectangular distribution has random variable X restricted to a finite interval $[a, b]$ and has $f(x)$ a constant over the interval. An illustration is shown in Figure 3:

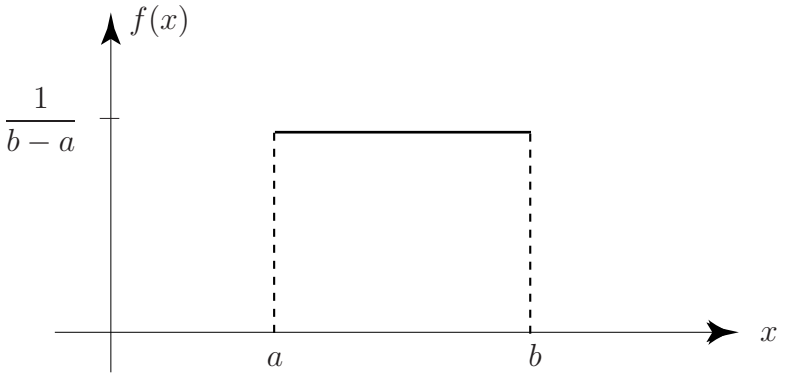


Figure 3

The function $f(x)$ is defined by:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

Mean and variance of a uniform distribution

Using the definitions of expectation and variance leads to the following calculations. As you might expect, for a uniform distribution, the calculations are not difficult.
Using the basic definition of expectation we may write:

$$\begin{aligned} \mathsf{E}(X) &= \int_{-\infty}^{\infty} x f(x)\, dx = \int_a^b x \frac{1}{b-a}\, dx = \frac{1}{2(b-a)} \left[x^2 \right]_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{b+a}{2} \end{aligned}$$

Using the formula for the variance, we may write:

$$\begin{aligned} \mathsf{V}(X) &= \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 \\ &= \int_a^b x^2 . \frac{1}{b-a}\, dx - \left(\frac{b+a}{2} \right)^2 = \frac{1}{3(b-a)} \left[x^3 \right]_a^b - \left(\frac{b+a}{2} \right)^2 \\ &= \frac{b^3 - a^3}{3(b-a)} - \left(\frac{b+a}{2} \right)^2 \\ &= \frac{b^2 + ab + a^2}{3} - \frac{b^2 + 2ab + a^2}{4} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

Key Point 3

The **Uniform** random variable X whose density function $f(x)$ is defined by

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

has expectation and variance given by the formulae

$$\mathrm{E}(X) = \frac{b+a}{2} \qquad \text{and} \qquad \mathrm{V}(X) = \frac{(b-a)^2}{12}$$

Example 2

The current (in mA) measured in a piece of copper wire is known to follow a uniform distribution over the interval $[0, 25]$. Write down the formula for the probability density function $f(x)$ of the random variable X representing the current. Calculate the mean and variance of the distribution and find the cumulative distribution function $F(x)$.

Solution

Over the interval $[0, 25]$ the probability density function $f(x)$ is given by the formula

$$f(x) = \begin{cases} \dfrac{1}{25-0} = 0.04, & 0 \le x \le 25 \\ 0 & \text{otherwise} \end{cases}$$

Using the formulae developed for the mean and variance gives

$$\mathrm{E}(X) = \frac{25+0}{2} = 12.5 \text{ mA} \qquad \text{and} \qquad \mathrm{V}(X) = \frac{(25-0)^2}{12} = 52.08 \text{ mA}^2$$

The cumulative distribution function is obtained by integrating the probability density function as shown below.

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

Hence, choosing the three distinct regions $x < 0$, $0 \le x \le 25$ and $x > 25$ in turn gives:

$$F(x) = \begin{cases} 0, & x < 0 \\ \dfrac{x}{25} & 0 \le x \le 25 \\ 1 & x > 25 \end{cases}$$

The thickness x of a protective coating applied to a conductor designed to work in corrosive conditions follows a uniform distribution over the interval $[20, 40]$ microns. Find the mean, standard deviation and cumulative distribution function of the thickness of the protective coating. Find also the probability that the coating is less than 35 microns thick.

Your solution

Answer

Over the interval $[20, 40]$ the probability density function $f(x)$ is given by the formula

$$f(x) = \begin{cases} 0.05, & 20 \leq x \leq 40 \\ 0 & \text{otherwise} \end{cases}$$

Using the formulae developed for the mean and variance gives

$$\mathrm{E}(X) = 10 \ \mu\text{m} \qquad \text{and} \qquad \sigma = \sqrt{\mathrm{V}(X)} = \frac{20}{\sqrt{12}} = 5.77 \ \mu\text{m}$$

The cumulative distribution function is given by

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

Hence, choosing appropriate ranges for x, the cumulative distribution function is obtained as:

$$F(x) = \begin{cases} 0, & x < 20 \\ \dfrac{x - 20}{20} & 20 \leq x \leq 40 \\ 1 & x \geq 40 \end{cases}$$

Hence the probability that the coating is less than 35 microns thick is

$$F(x < 35) = \frac{35 - 20}{20} = 0.75$$

Exercises

1. In the manufacture of petroleum the distilling temperature (T°C) is crucial in determining the quality of the final product. T can be considered as a random variable uniformly distributed over 150°C to 300°C. It costs $£C_1$ to produce 1 gallon of petroleum. If the oil distills at temperatures less than 200°C the product sells for $£C_2$ per gallon. If it distills at a temperature greater than 200°C it sells for $£C_3$ per gallon. Find the expected net profit per gallon.

2. Packages have a nominal net weight of 1 kg. However their actual net weights have a uniform distribution over the interval 980 g to 1030 g.

   (a) Find the probability that the net weight of a package is less than 1 kg.

   (b) Find the probability that the net weight of a package is less than w g, where $980 < w < 1030$.

   (c) If the net weights of packages are independent, find the probability that, in a sample of five packages, all five net weights are less than wg and hence find the probability density function of the weight of the heaviest of the packages. (Hint: all five packages weigh less than w g if and only if the heaviest weighs less that w g).

Answers

1.

$$\mathsf{P}(X < 200) = 50 \times \frac{1}{150} = \frac{1}{3} \qquad \mathsf{P}(X > 200) = \frac{2}{3}$$

Let F be a random variable defining profit.

F can take two values $£(C_2 - C_1)$ or $£(C_3 - C_1)$

x	$C_2 - C_1$	$C_3 - C_1$
$\mathsf{P}(F = x)$	$^1/_3$	$^2/_3$

$$\mathsf{E}(F) = \left[\frac{C_2 - C_1}{3}\right] + \frac{2}{3}[C_3 - C_1] = \frac{C_2 - 3C_1 + 2C_3}{3}$$

2.

(a) The required probability is $\mathsf{P}(W < 1000) = \dfrac{1000 - 98}{1030 - 980} = \dfrac{20}{50} = 0.4$

(b) The required probability is $\mathsf{P}(W < w) = \dfrac{w - 980}{1030 - 980} = \dfrac{w - 980}{50}$

(c) The probability that all five weigh less than w g is $\left(\dfrac{w - 980}{50}\right)^5$ so the pdf of the heaviest is

$$\frac{d}{dw}\left(\frac{w - 980}{50}\right)^5 = \frac{5}{50}\left(\frac{w - 980}{50}\right)^4 = 0.1\left(\frac{w - 980}{50}\right)^4 \quad \text{for } 980 < w < 1030.$$

1. The exponential distribution

The exponential distribution is defined by

$$f(t) = \lambda \mathrm{e}^{-\lambda t} \qquad t \geq 0 \qquad \lambda \text{ a constant}$$

or sometimes (see the Section on Reliability in HELM 46) by

$$f(t) = \frac{1}{\mu}\mathrm{e}^{-t/\mu} \qquad t \geq 0 \qquad \mu \text{ a constant}$$

The advantage of this latter representation is that it may be shown that the mean of the distribution is μ.

Example 3

The lifetime T (years) of an electronic component is a continuous random variable with a probability density function given by

$$f(t) = \mathrm{e}^{-t} \qquad t \geq 0 \qquad \text{(i.e. } \lambda = 1 \text{ or } \mu = 1\text{)}$$

Find the lifetime L which a typical component is 60% certain to exceed. If five components are sold to a manufacturer, find the probability that at least one of them will have a lifetime less than L years.

Solution

We require $\mathrm{P}(T > L) = 0.6$. We know that this probability is given by the relationship

$$\mathrm{P}(T > L) = \int_L^\infty \mathrm{e}^{-t}\,dt = \Big[-\mathrm{e}^{-t} \Big]_L^\infty = \mathrm{e}^{-L}$$

Solving $\mathrm{e}^{-L} = 0.6$ for the least value of L we obtain $L = 0.51$ years.
Assuming that the lifetime of each component is independent we have

P(at least one component has a lifetime less than 0.51 years)

$= 1 -$ P(no component has a lifetime less than 0.51 years)

$= 1 - 0.6^5$

$= 0.92$

Commonly, car cooling systems are controlled by electrically driven fans. Assuming that the lifetime T in hours of a particular make of fan can be modelled by an exponential distribution with $\lambda = 0.0003$ find the proportion of fans which will give at least 10000 hours service. If the fan is redesigned so that its lifetime may be modelled by an exponential distribution with $\lambda = 0.00035$, would you expect more fans or fewer to give at least 10000 hours service?

Your solution

Answer

We know that $f(t) = 0.0003\mathrm{e}^{-0.0003t}$ so that the probability that a fan will give at least 10000 hours service is given by the expression

$$\mathrm{P}(T > 10000) = \int_{10000}^{\infty} f(t)\,dt = \int_{10000}^{\infty} 0.0003\mathrm{e}^{-0.0003t}\,dt = -\left[\mathrm{e}^{-0.0003t}\right]_{10000}^{\infty} = e^{-3} \approx 0.0498$$

Hence about 5% of the fans may be expected to give at least 10000 hours service. After the redesign, the calculation becomes

$$\mathrm{P}(T > 10000) = \int_{10000}^{\infty} f(t)\,dt = \int_{10000}^{\infty} 0.00035\mathrm{e}^{-0.00035t}\,dt = -\left[\mathrm{e}^{-0.00035t}\right]_{10000}^{\infty} = e^{-3.5} \approx 0.0302$$

and so only about 3% of the fans may be expected to give at least 10000 hours service.

Hence, after the redesign we expect *fewer* fans to give 10000 hours service.

Exercises

1. The time intervals between successive barges passing a certain point on a busy waterway have an exponential distribution with mean 8 minutes.

   (a) Find the probability that the time interval between two successive barges is less than 5 minutes.

   (b) Find a time interval t such that we can be 95% sure that the time interval between two successive barges will be greater than t.

2. It is believed that the time X for a worker to complete a certain task has probability density function $f_X(x)$ where

$$f_X(x) = \begin{cases} 0 & (x \leq 0) \\ kx^2 e^{-\lambda x} & (x > 0) \end{cases}$$

   where λ is a parameter, the value of which is unknown, and k is a constant which depends on λ.

   (a) Show that if $\quad I_n = \displaystyle\int_0^\infty x^n e^{-\lambda x}\, dx \quad$ then $\quad I_n = \dfrac{n}{\lambda} I_{n-1}, \quad$ where $n > 0$ and $\lambda > 0$.

   Evaluate $\quad I_0 = \displaystyle\int_0^\infty e^{-\lambda x}\, dx \quad$ and hence find a general expression for I_n.

   This result can be used in the rest of this question.

   (b) Find, in terms of λ, the value of k.

   (c) Find, in terms of λ, the expected value of X.

   (d) Find, in terms of λ, the variance of X.

   (e) Write down the expected value and variance of the sample mean of a sample of n independent observations on X.

   (f) Find, in terms of λ, the expected value of X^{-1}.

Answers

1. We have $\mu = 8$ so $\lambda = 0.125$.

   (a) The probability is

   $$\mathsf{P}(T < 5) = \int_0^5 0.125e^{-0.125t}\,dt = 1 - e^{-0.125\times 5} = 0.4647.$$

   (b) We require

   $$\int_t^\infty 0.125e^{-0.125x}\,dx = e^{-0.125t} = 0.95.$$

   So $-0.125t = \log 0.95$ and

   $$t = -\frac{\log 0.95}{0.125} = 0.4103.$$

   That is, 24.6 s.

2.

(a) $I_n = \displaystyle\int_0^\infty x^n e^{-\lambda x}\,dx = \left[-\frac{1}{\lambda}x^n e^{-\lambda x}\right]_0^\infty + \frac{n}{\lambda} \qquad \int_0^\infty x^{n-1}e^{-\lambda}\,dx = \frac{n}{\lambda}I_{n-1}$

$I_0 = \displaystyle\int_0^\infty e^{-\lambda x}\,dx = \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda}$ hence $I_n = \dfrac{n!}{\lambda^{n+1}}$.

(b) $\displaystyle\int_0^\infty kx^2e^{-\lambda x}\,dx = 1 \Rightarrow kI_2 = 1 \Rightarrow k = \frac{1}{I_2} = \frac{\lambda^3}{2}$

(c) $\mathsf{E}(X) = \displaystyle\int_0^\infty x f_X(x)\,dx = kI_3 = \frac{\lambda^3}{2}\frac{6}{\lambda^4} = \frac{3}{\lambda}$

(d) $\mathsf{E}(X^2) = \displaystyle\int_0^\infty x^2 f_X(x)\,dx = kI_4 = \frac{\lambda^3}{2}\frac{24}{\lambda^5} = \frac{12}{\lambda^2}$

so $\mathsf{V}(X) = \mathsf{E}(X^2) - \{\mathsf{E}(X)\}^2 = \dfrac{12}{\lambda^2} - \dfrac{9}{\lambda^2} = \dfrac{3}{\lambda^2}$

(e) $\mathsf{E}(\bar{X}) = \dfrac{3}{\lambda} \qquad \mathsf{V}(\bar{X}) = \dfrac{3}{n\lambda^2}$

(f) $\mathsf{E}\left(\dfrac{1}{X}\right) = \displaystyle\int_0^\infty \frac{1}{x} f_X(x)\,dx - kI_1 = \frac{\lambda^3}{2}\frac{1}{\lambda^2} = \frac{\lambda}{2}$

11

CORRELATION AND REGRESSION

CHAPTER OUTLINES

• Introduction • Correlation • Scatter or Dot Diagram • Limitation of Correlation • Karl Pearson Coefficient of Correlation • Rank Correlation Coefficient • Limitation of r_{xy} • Properties of Correlation Coefficient r_{xy} • Regression • Curve Fitting • Fitting of a Straight Line • Multiple Regression • Curvilinear Regression • Multiple and Partial Correlation • Residual • Variance of a Residual.

11.1 INTRODUCTION

We know that the volume V of a cube of side a is given by $V = a^3$. This gives us that a cube with larger sides will always have a larger volume than a cube with a smaller sides. Hence there exists a fundamental relationship between V and a. Now if we consider the heights and weights of students in a class. We cannot put the statement "A tall student is more likely to be heavier weight than a short student."

In this chapter we shall study the relationship of the type between height and weight or height and age or price and demand etc. Such types of a relationship is **called statistical relationship**. Special methods have been developed to discover the existance of statistical relationship, between the two variables from the bivariate data, if both the variables in bivariate data are quantitative, we use the term **Correlation Analysis**. To describe the methods designed to find out if the statistical relationship between the two variables exists or not, but in a bivariate data having both the variables as quantiative variables, the methods of estimating the likely value of one variable from the known value of the other variable form a part of **Regression Analysis**. According to the statistical **A.M. Tuttle, "correlation is an analysis of the covariation between two or more variables."**

11.2 CORRELATION

Here we shall study the relationship between two variables (Independent or variables) such that a change in one is accompanied by change in the other in such a way that an increase in one is accompanied by an increase or decrease in the other is called a correlation.

Many Engineering and scientific problems are concerned with determining a relationship between a set of variables. The number of criminals and number of policemen, the number of ticket windows and number of passengers... etc. in all such cases to analyse the strength of the relationship between two variables.

11.3 TYPES OF CORRELATIONS

(i) Positive Correlation

If an increase or decrease in the values of one variable is always corresponding and proportional increase or decrease in other variables. Then the corrrelation will be perfect positive correlation. For example, age of mothers and children are known to have a positive correlation.

(ii) Negative Correlation

If an increase or decrease in the values of one variable is always corresponding and proportional decrease or increase in other variable then the correlation will be perfect negative correlation, for example, number of faculty members and number of staff rooms in college, demand and supply are known to have a negative correlation.

(iii) Linear Correlation

If the two variables (x or y) are represented by a line i.e. represented by line ($y = a + bx$) then the correlation is said to be linear correlation.

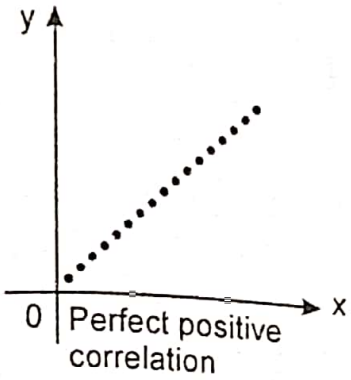
(iv) Non Linear Correlation

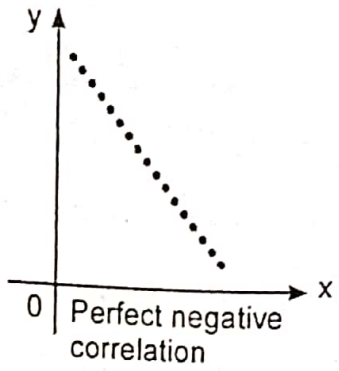
If the two variables (x and y) are not represented by a line i.e. represented by an equation of the form $y = a + bx + cx^2$, the correlation is said to be a non linear correlation.

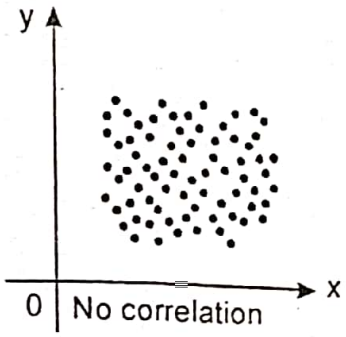
11.4 SCATTER OR DOT DIAGRAM

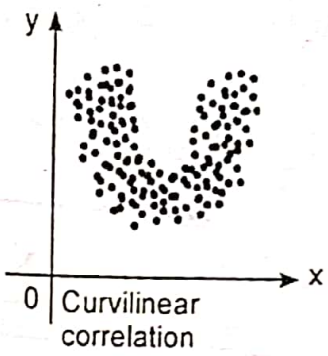
Scatter diagram is the simplest way of the diagrammatic representation of bivariate data. Let (x_i, y_i) ; $i = 1, 2, ...n$ be a bivariate distribution. If the values of the variables x and y be plotted along the x-axis and y-axis respectively in the x-y plane. The diagram of dots so obtained is known as **Scatter diagram** or **dot diagram.**

The scatter diagram may indicate both degree and the type of correlation. From scatter diagram we can form a fairly good, though rough, idea about the relationship between the two variables. The different types of correlation are depicted by means of scattered diagrams as shown below:









If in scatter diagram, the points are very dense means very close to each other, we should expect a fairly good amount of correlation between the variable x and y. And if the points are widely scattered, a poor correlation is expected.

Advantages and Disadvantages of Scatter Diagram

Advantages :

1. It is readily comprehensible and enables us to form a rough idea of the nature of relationship. Between the variables.
2. It is not affected by extreme observations.
3. It is not influenced by the size of extreme items.

Disadvantages :

1. It is not suitable if the number of observations is very large.
2. It is enables us to obtain an approximate estimating line or line of best fit by this method.
3. It is ony a rough measure of correlation where the exact magnitude cannot be known.

11.5 KARL PEARSON COEFFICIENT OF CORRELATION

Karl Pearson has given a coefficient to measure the degree of linear relationship between two variables, which is known as coefficient of correlation. For bivariate distribution (x, y) the coefficient of correlation denoted by r_{xy} and is denoted as

$$r_{xy} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)}.\sqrt{\operatorname{var}(y)}} = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} \qquad \ldots(1)$$

Where $\operatorname{cov}(x,y)$ = covariance of x and y

σ_x = variance of x

σ_y = variance of y

$$\operatorname{cov}(x,y) = E\left[(x-\bar{x})(y-\bar{y})\right]$$

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

Here $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

and $\sigma_x^2 = \operatorname{var}(x) = E\left[(x-\bar{x})^2\right]$

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

Similarly $\sigma_y^2 = \operatorname{var}(y)$

$$= E\left[(y-\bar{y})^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

Substitute the values of $\operatorname{cov}(x,y)$, σ_x and σ_y in equation (1) we get

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$\boxed{r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}} \qquad \ldots(2)$$

$cov.(x,y)$ can be simplify as (2)

$$\text{cov}\cdot(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\sum_{i=1}^{n}\frac{y_i}{n} - \bar{y}\sum_{i=1}^{n}\frac{x_i}{n} + \bar{x}\,\bar{y}$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\,\bar{y} - \bar{y}\,\bar{x} + \bar{x}\,\bar{y}$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\,\bar{y}$$

$\gamma_{xy} = \dfrac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$

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

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

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

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

Similarly $$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}y_i^2 - \bar{y}^2$$

Substituting the above values in (2), we get

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2}\sqrt{\frac{1}{n}\sum_{i-1}^{n}y_i^2 - \bar{y}^2}}$$

Note : *We can write* $\text{cov}(x,y) = \sigma_{xy} = \mu_{11}$, *i.e. Karl Pearson's coefficient of correlation is also called product moment correlation coefficient.*

Remark : *The two most popular correlation coefficients are Spearman's correlation coefficient and Pearson's product moment correlation coefficient, when calculating a correlation coefficient for ordinal data, select Spearman's technique and for internal or ratio-type data, use Pearson's technique..*

11.6 LIMITATION OF R_{XY}

(i) The coefficient of correlation can be used as a measure of linear relationship between two variables. In case of non linear or any other relationship the coefficient of correlation does not provide any measure at all. So the inspection of Scatter diagram is essential.

(ii) Correlation must be used to the data drawn from the same source. If distinct sources are used then the two variables may show correlation but in each source they may be uncorrelated.

(iii) If positive or negative correlation exists between two variables then it may be due to the effect of some other variables in both of them on the elimination of this effect. It may be found that the net correlation is nil.

11.7 PROPERTIES OF CORRELATION COEFFICIENT R_{XY}

(i) Coefficient of correlation r_{xy} is independent change of scale and origin.

(ii) $-1 \le r_{xy} \le 1$ i.e. the coefficient of correlation r_{xy} lies between -1 to 1.

[Raj. Univ. CS 2003]

Proof (i): Suppose $u = \dfrac{x-a}{h}, \quad v = \dfrac{y-b}{k},$

So that $x = a + hu$ and $y = b + vk$ where a, b, h, k are constants, where $h > 0, k > 0$.

We will prove that $r(x, y) = r(u, v)$

$\because \quad x = a + h$ and $y = b + u_k$

$\therefore \quad E(x) = a + hE(u)$ and $E(y) = b + kE(v)$

$\Rightarrow \quad x - E(x) = h[u - E(u)]$

and $\quad y - E(y) = k[v - E(v)]$

Therefore $\quad \text{cov}\cdot(x,y) = E[\{x - E(x)\}\{y - E(y)\}]$

$$= E\left[h\{u-E(u)\}k\{v-E(v)\}\right]$$

$$= hk\,E\left[\{u-E(u)\}\{V-E(v)\}\right]$$

$$\text{cov}\cdot(x,y) = hk\,\text{cov}\cdot(u,v) \qquad \ldots(3)$$

$\therefore$

and $$\sigma_x^2 = E\left[\{x-E(x)\}^2\right] = E\left[h^2\{i-E(u)\}\right]$$

$\Rightarrow$ $$\sigma_x^2 = h^2\,\sigma_y^2$$

or $$\sigma_x = h\sigma_y\ ;\ h>0 \qquad \ldots(4)$$

and $$\sigma_y^2 = E\left[\{y-E(y)\}^2\right] = E\left[k^2\{V-E(v)\}^2\right]$$

$\Rightarrow$ $$\sigma_y^2 = k^2\sigma_v^2$$

or $$\sigma_y = k\,\sigma_v\ ;\ k>0 \qquad \ldots(5)$$

Substituting from (3), (4) and (5) in (1) ;

We get $$r(x,y) = \frac{\text{cov}\cdot(x,y)}{\sigma_x\sigma_y} = \frac{hk\,\text{cov}\cdot(u,v)}{hk\,\sigma_u\,\sigma_v}$$

$$= \frac{\text{cov}.(u,v)}{\sigma_u\ \sigma_v} = r(u,v)$$

This theorem is of fundamental importance inthe numerical computation of the coefficient.

Proof (ii): Let $u_i = \dfrac{x_i-\bar{x}}{\sigma_x}$ and $V_i = \dfrac{y_i-\bar{y}}{\sigma_y}$

$$\frac{1}{n}\sum u_i^2 = \frac{1}{n}\,\frac{\Sigma(x_i-\bar{x})^2}{\sigma_x^2}$$

$$= \frac{1}{n}\,\frac{\Sigma(x_i-\bar{x})^2}{\dfrac{\Sigma(x_i-\bar{x})^2}{n}} = 1$$

$\Rightarrow$ $$\frac{1}{n}\sum u_i^2 = 1$$

Similarly $\frac{1}{n}\sum v_i^2 = 1$

and $\frac{1}{n}\sum u_i v_i = \frac{1}{n}\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x \sigma_y} = r_{xy}$

Now $\frac{1}{n}\sum(u_i - v_i)^2 \geq 0$

$\frac{1}{n}\sum u_i^2 - \frac{2}{n}\sum u_i v_i + \frac{1}{n}\sum v_i^2 \geq 0$

$1 - 2r_{xy} + 1 \geq 0$

$2(1 - r_{xy}) \geq 0 \Rightarrow 1 - r_{xy} \geq 0$

$1 \geq r_{xy} \Rightarrow r_{xy} \leq 1$ ...(6)

and $\frac{1}{n}\sum (u_i + v_i)^2 \geq 0$

$\frac{1}{n}\sum u_i^2 + \frac{2}{n}\sum u_i v_i + \frac{1}{n}\sum v_i^2 \geq 0$

$1 + 2r_{xy} + 1 \geq 0$

$\Rightarrow 2(1 + r_{xy}) \geq 0 \Rightarrow 1 + r_{xy} \geq 0$

$\Rightarrow r_{xy} \geq -1$ ...(7)

$\boxed{-1 \leq r_{xy} \leq 1}$

(iii) Two independent variables are perfect negative linear correlation if $r_{xy} = -1$

(iv) Two independent variables are perfect positive linear correlation if $r_{xy} = 1$

(v) Two independent variables are uncorrelated if $r_{xy} = 0$ but the converse is not always true, i.e. A correlation coefficient is zero means there is no relationship between the two variables.

(vi) If $r_{xy} = r_{yx}$, r_{xy} is called product moment correlation coefficient by summetry.

Example 1. Calculate the coefficient of correlation between x and y using the following data.

x	1	3	5	7	8	10
y	8	12	15	17	18	20

[Raj. Univ. CS 2006]

Solution :

x	y	$x-\bar{x}=x-6$	$y-\bar{y}=y-15$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	8	−5	−7	25	49	35
3	12	−3	−3	9	9	9
5	15	−1	0	1	0	0
7	17	1	2	1	4	2
8	18	2	3	4	9	6
10	20	4	5	16	25	20
$\Sigma x=36$	$\Sigma y=90$	$\Sigma(x-\bar{x})=-2$	$\Sigma(y-\bar{y})=0$	$\Sigma(x-\bar{x})^2=56$	$\Sigma(y-\bar{y})^2=96$	$\Sigma(x-\bar{x})(y-\bar{y})=72$

Here $n = 6$

then $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{36}{6} = 6$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{90}{6} = 15$

$$r_{xy} = \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y-\bar{y})^2}}$$

$$= \frac{\frac{72}{6}}{\sqrt{\frac{56}{6}}\sqrt{\frac{96}{6}}} = \frac{72}{73.32}$$

$\boxed{r_{xy} = 0.98}$ **Ans.**

Example 2. **Calculate the coefficient of correlation for x and y from the following data.**

x	45	55	56	58	60	65	68	70	75	80	85
y	56	50	48	60	62	64	65	70	74	82	90

[Raj. Univ. CS 2005]

Solution:

x	y	x^2	y^2	xy
45	56	2025	3136	2520
55	50	3025	2500	2750
56	48	3136	2304	2688
58	60	3364	3600	3480
60	62	3600	3844	3720
65	64	4225	4096	4160
68	65	4624	4225	4420
70	70	4900	4900	4900
75	74	5625	5476	5550
80	82	6400	6724	6560
85	90	7225	8100	7650
$\sum x = 717$	$\sum y = 721$	$\sum x^2 = 48149$	$\sum y^2 = 48905$	$\sum xy = 48398$

Karl Pearson correlation coefficient

$$r_{xy} = \frac{\text{cov}(xy)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}\ \sqrt{\frac{1}{n}\sum y_i^2 - \bar{y}^2}}$$

here $n = 11$ $\quad \bar{x} = \dfrac{\sum x_i}{n} = \dfrac{717}{11} = 65.181$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{721}{11} = 65.545$$

$$\therefore \quad r_{xy} = \frac{\frac{48398}{11} - 65.181 \times 65.545}{\sqrt{\frac{48149}{11} - (65.181)^2}\ \sqrt{\frac{48905}{11} - (65.545)^2}}$$

$$= \frac{4399.818 - 4272.318}{\sqrt{4377.181 - 4248.56}\ \sqrt{4445.909 - 4296.147}}$$

$$r_{xy} = \frac{127.5}{\sqrt{128.618}\ \sqrt{149.761}}$$

$$= \frac{127.5}{138.775} = 0.918$$

$\boxed{r_{xy} = 0.92}$ **Ans.**

Example 3. Calculate the coefficient of correlation between x and y from the following data

x	−10	−5	0	5	10
y	5	9	7	11	13

[RTU, B.Tech. 2008]

Solution:

x	y	x^2	y^2	xy
−10	5	100	25	−50
−5	9	25	81	−45
5	7	0	49	0
5	11	25	121	5
10	13	100	169	130
$\Sigma x = 0$	$\Sigma y = 45$	$\Sigma x^2 = 250$	$\Sigma y^2 = 445$	$\Sigma xy = 40$

Correlation coefficient

$$r_{xy} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}\ \sqrt{\frac{1}{n}\sum y_i^2 - \bar{y}^2}}$$

Here $n = 5$, $\quad \bar{x} = \frac{\Sigma x_i}{n} = 0$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{45}{5} = 9$$

$$\therefore \quad r_{xy} = \frac{\frac{40}{5} - 0 \times 9}{\sqrt{\frac{250}{5} - 0}\ \sqrt{\frac{445}{5} - 81}} = \frac{8}{\sqrt{50}\sqrt{8}}$$

$$= \frac{8}{7.071 \times 2.828} = \frac{8}{19.999}$$

$\boxed{r_{xy} = 0.040}$ **Ans.**

11.8 RANK CORRELATION COEFFICIENT

This method is finding out covariability or the lack of it between two variables it was developed by the British psychologist **Charles Edward Spearman** in 1904.

This measure is especially useful when quantitative measures for certain factors (such as the judgement of female beauty or in the evaluation of leadership in the group) can not be fixed, the individual group can be arranged in order there by obtaining for each individual a number indicating his or her rank in the group. Let (x_i, y_i), $i = 1, 2, 3, ...,n$ be the ranks of the ith individual in two characteristics A and B respectively. Pearsonian coefficient of correlation between the ranks x'_is and y'_is is called the rank correlation coefficient between A and B for that group of individuals.

11.8.1 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coeffieient is defined as

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

where ρ denotes rank coeffieient of correlation and d_i refers to the difference of rank between paired items in two series.

Consider a set of n individuals, and also assuming that no two individuals are bracketed equal in either classification, each of the variables x and y takes the values 1, 2, 3, ...n.

Therefore $$\bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + 3 + ...n) = \frac{n(n+1)}{2n}$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\left(1^2 + 2^2 + 3^2 + ... + n^2\right) - \left(\frac{n+1}{2}\right)^2$$

$$\sigma_x^2 = \frac{1}{n}\frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2 - 1}{12}$$

$$\Rightarrow \quad \sigma_x^2 = \frac{n^2 - 1}{12} = \sigma_y^2$$

In general

$x_i \neq y_i$, suppose $d_i = x_i - y_i$

$$\therefore \quad d_i = (x_i - \bar{x}) - (y_i - \bar{y})$$

$$\Rightarrow \quad \sum d_i^2 = \sum_i \left[(x_i - \bar{x}) - (y_i - \bar{y})\right]^2$$

$$= \sum_i (x_i - \bar{x})^2 - 2\sum (x_i - \bar{x})(y_i - \bar{y}) + \sum (y_i - \bar{y})^2$$

Dividing both the sides by n, we get

$$\frac{1}{n}\sum d_i^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 - \frac{2}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) + \frac{1}{n}\sum (y_i - \bar{y})^2$$

$$\frac{1}{n}\sum d_i^2 = \sigma_x^2 + \sigma_y^2 - 2\,\text{cov}\cdot(x, y)$$

$$= \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$$

$$= \sigma_x^2 + 2\rho\sigma_x^2 \qquad \left[\because \sigma_x^2 = \sigma_y^2\right]$$

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

Which is the Spearman's formula for the rank correlation coefficient.

Remark : *Spearman's Rank correlation coefficient is also noted that* $-1 < r_{xy} < 1$ *and the above formula is used when ranks are not repeated.*

11.8.2 Rank Correlation Coefficient for Repeated Ranks

If some of the individuals receive the same rank in ranking of merit, they are said to be tied.

If m is the number of times an item is repeated then the factor $\frac{m\left(m^2 - 1\right)}{12}$ is to be added to Σd^2. This correlation factor is to be added for each repeated rank, which is represented by the formula.

$$r = 1 - \frac{6\left\{\Sigma d_i^2 + \frac{1}{12}m\left(m^2 - 1\right)\right\}}{n\left(n^2 - 1\right)}$$

Example 5. The marks secured by students in mathematics and statistics are given below:

Mathematics (x)	10	15	12	17	13	16	25	14	22
Statistics (y)	30	42	45	46	33	34	40	35	39

Solution :

Math (x)	Statistics (y)	Rank of x R_1	Rank of y R_2	$d = R_1 - R_2$	d^2
10	30	9	9	0	0
15	42	5	3	2	4
12	45	8	2	6	36
17	46	3	1	2	4
13	33	7	8	−1	1
16	34	4	7	−3	9
25	40	1	4	−3	9
14	35	6	6	0	0
22	39	2	5	−3	9
				$\Sigma d = 0$	$\Sigma d^2 = 72$

$\therefore$ Rank correlation coefficient

$$r = 1 - \frac{6\Sigma d_i^2}{n\left(n^2 - 1\right)}$$

Here $n = 9$, $\sum d_i^2 = 72$

$$r = 1 - \frac{6 \times 72}{9 \times 80} = 1 - \frac{432}{720}$$

$\boxed{r = 0.4}$ **Ans.**

Example 5. Obtain the rank correlation coefficient of the following data

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

Solution:

x	y	Rank of x d_1	Rank of y d_2	$d = d_1 - d_2$	d^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	6	10	-1	1
64	81	9	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				$\Sigma d = 0$	$\Sigma d^2 = 72$

Since ranks are repeated then rank correlation coefficient.

$$r = \frac{1-6\left\{\Sigma d_i^2 + \frac{m(m^2-1)}{12}\right\}}{n(n^2-1)} = \frac{1-6\left\{\Sigma d_i^2 + F\right\}}{n(n^2-1)}$$

First we calculated F for repeated rank (separately for x and y)

In x-term 75 repeated twice ($m_1 = 2$)

64 repeated thrice ($m_2 = 3$)

then $$F_1 = \frac{m_1\left(m_1^2-1\right)}{12} + \frac{m_2\left(m_2^2-1\right)}{12}$$

$$F_1 = \frac{2(4-1)}{12} + \frac{3(9-1)}{12} = \frac{1}{2} + 2 = \frac{5}{2}$$

in y-term, 68 repeated twice ($m = 2$)

$$F_2 = \frac{m\left(m^2-1\right)}{12} = \frac{2(4-1)}{12} = \frac{1}{2}$$

$$F_2 = \frac{1}{2}$$

Then $F = F_1 + F_2$

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

Now Rank of correlation coefficient

$$r = 1 - \frac{6(72 + 0.3)}{10(100 - 1)} = 1 - \frac{6 \times 75}{10 \times 99} = 1 - 0.4545$$

$\boxed{r = 0.545}$ **Ans.**

Example 6. Ten competitors in beauty contest are ranked by three judges in the following order.

First Judge (R_1)	1	6	5	10	3	2	4	9	7	8
Second Judge (R_2)	3	5	8	4	7	10	2	1	6	9
Third Judge (R_3)	6	4	9	8	1	2	3	10	5	7

use the rank correlation method and discuss which pair of judges have the nearest approach to common testes in beauty.

Solution:

R_1	R_2	R_3	R_1-R_2 = d_1	R_2-R_3 = d_2	R_1-R_3 = d_3	d_1^2	d_2^2	d_3^2
1	3	6	–2	–3	–5	4	9	25
6	5	4	1	1	2	1	1	4
5	8	9	–3	–1	–4	9	1	16
10	4	8	6	–4	2	36	16	4
3	7	1	–4	6	2	16	36	4
2	10	2	–8	8	0	64	64	0
4	2	3	2	–1	1	4	1	1
9	1	10	8	–9	–1	64	81	1
7	6	5	1	1	2	1	1	4
8	9	7	–1	–2	–1	1	4	1
			$\Sigma d_1 = 0$	$\Sigma d_2 = 0$	$\Sigma d_3 = 0$	$\Sigma d_1^2 = 200$	$\Sigma d_2^2 = 214$	$\Sigma d_3^2 = 60$

Here $n = 10$

Rank correlation coefficient for R_1 and R_2

$$r = 1 - \frac{6\Sigma d_1^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 200}{10\times 99} = \frac{-7}{33}$$

$$r = -0.212$$

For Rank R_2 and Rank R_3

$$r = 1 - \frac{6\Sigma d_2^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 214}{10\times 99} = \frac{-49}{165}$$

$$r = -0.297$$

For Rank R_1 and Rank R_3

$$r = 1 - \frac{6\Sigma d_3^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 60}{10\times 99} = \frac{7}{11}$$

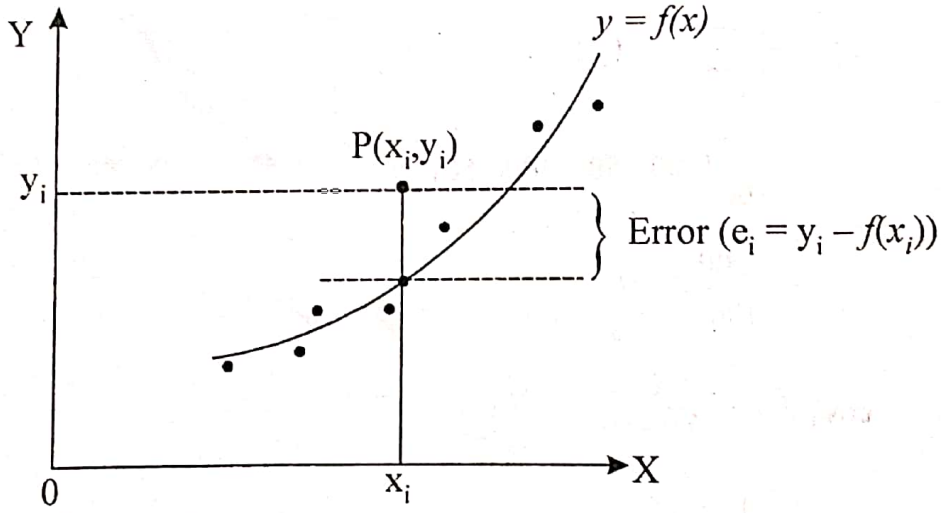
$$r = 0.636$$

We see that the first and third judges have the nearest approach to their tastes for beauty.

11.10 METHOD OF LEAST SQUARES

Generally a mathematical equation is fitted to experimental data by plotting the data on a graph paper and then passing a straight line through the data points. The drawback of this method is that the straight line drawn may not be unique. The method of least squares is probably the most systematic procedures to fit a unique curve through given data points and it is used in modern computer.

Let us suppose that x and y be two variables which give us a set of n pairs of numerical values $(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)$. The scatter diagram for given data is



Scatter diagram

Let us suppose that $y = f(x)$ be an approximation to the function.

The deviations (errors e_i's) between obtain from the approximations of y's and the true value (actual tabulated) of y's are

$$e_i = y_i - f(x_i)$$

for $i = 1,2,3,....,n$ ...(1)

So the **least square method states that a curve is a best fit, if the sum of the squares of deviations of the individual points from the curve is minimum.**

i.e. we form the sum S(say) the squares of deviations, as

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - f(x_i) \right]^2 \qquad \ldots(2)$$

If S = 0 then $y_i = f(x_i)$ for all i, therefore all the points lie on the curve otherwise the sum will be either minimum or maximum, if the partial derivatives of type $\frac{\partial S}{\partial a} = 0$, where a is one of the unkowns assumed in the approximation $y = f(x)$.

In the following sections, we shall study the linear and nonlinear least squares fitting to given data $(x_i, y_i), i = 1, 2, \ldots, n$.

11.10.1 Fitting a Straight Line

Let us suppose that $y = a + bx$ be the straight line to be fitted to the given data (x_i, y_i), where a and b represent intercept of the line and slope respectively, then corresponding to equation (2), we have

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - f(x_i) \right]^2$$

$$= \left[y_1 - f(x_1) \right]^2 + \left[y_2 - f(x_2) \right]^2 + \ldots + \left[y_n - f(x_n) \right]^2$$

$$= \left[y_1 - (a + bx_1) \right]^2 + \left[y_2 - (a + bx_2) \right]^2 + \ldots + \left[y_n - (a + bx_n) \right]^2 \qquad \ldots(3)$$

By the principle of least squares, S is minimum

$$\therefore \quad \frac{\partial S}{\partial a} = 0 \text{ and } \frac{\partial S}{\partial b} = 0$$

$$\therefore \quad 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0 \qquad \ldots(4)$$

and $$2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0 \qquad \ldots(5)$$

After simplication, we get the following linear equations

$$na + b\Sigma x = \Sigma y \qquad \ldots(6)$$

$$a\Sigma x + b\Sigma x^2 = \Sigma xy \qquad \ldots(7)$$

These equations are called **"NORMAL EQUATIONS"** for the least square line, since x_i and y_i are known, we can solve these equations by Cramer's rule for a and b.

where $b = \dfrac{\begin{vmatrix} n & \Sigma y \\ \Sigma x & \Sigma xy \end{vmatrix}}{\begin{vmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{vmatrix}}$ and $a = \dfrac{\begin{vmatrix} \Sigma y & \Sigma x \\ \Sigma xy & \Sigma x^2 \end{vmatrix}}{\begin{vmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{vmatrix}}$

Now differentiating equation (4) and (5) with respect to a an b respectively, we find that $\dfrac{\partial^2 S}{\partial a^2}$ and $\dfrac{\partial^2 S}{\partial b^2}$ will both be positive at the point a and b. Therefore these values provide a minimum of S.

Again, dividing equation (4) throughout by n, we obtain

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

or $$a + b\bar{x} = \bar{y}$$

Where $(\bar{x}, \bar{y})$ is the controid of the given data points, it means, the fitted straight line passes through the centroid of the data points.

Note: *In case of change of origin means when the numbers in the given data are large, we can change the origin and scale with help of substitutions.*

$$X = \frac{x - (middle\,term)}{interval}, \text{ when n is odd.}$$

or $$X = \frac{x - (\text{mean of two middle terms})}{\frac{1}{2}(\text{interval})}, \text{ when } n \text{ is even.}$$

Example 8. Fit a straight line to the following data :

x:	0	1	2	3	4
y:	1	1.8	3.3	4.5	6.3

Solution : Let the straight line of best fit be

$$y = a + bx$$

The normal equations are

$$\Sigma y = na + b\Sigma x \quad \ldots(1)$$

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

$$\ldots..(3)$$

Here $n = 5$

The values of $\Sigma x, \Sigma y, \Sigma xy$ and Σx^2 can be computed as in the following table

x	y	x^2	xy
0	1	0	1
1	1.8	1	1.8
2	3.3	4	6.6
3	4.5	9	13.5
4	6.3	16	25.2
$\Sigma x = 10$	$\Sigma y = 16.9$	$\Sigma x^2 = 30$	$\Sigma xy = 47.1$

From (2) and (3)

$$16.9 = 5a + 10b$$

and $$47.1 = 10a + 30b$$

Solving the equations, we get

$$a = 0.72 \text{ and } b = 1.33$$

Hence, the required straight line is

$$y = 0.72 + 1.33x \quad \textbf{Ans.}$$

11.10.2 Fitting a Parabola (Non-Linear)

Let us suppose that $y = a + bx + cx^2$ be a parabola to be fitted to the given data $(x_i, y_i), i = 1,2,3,...,n$

The deviation (error) at $x = x_i$ is

$$e_i = y_i - f(x_i)$$

$$e_i = y_i - (a + bx_i + cx_i^2) \quad \text{....(1)}$$

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (a + bx_i + cx_i^2) \right]^2 \quad \text{....(2)}$$

So, by principle of least squares, S should be minimum for the best values of a, b and c.

$$\therefore \quad \frac{\partial S}{\partial a}=0, \frac{\partial S}{\partial b}=0 \text{ and } \frac{\partial S}{\partial c}=0$$

or
$$\sum_{i=1}^{n} 2\left[y_i-\left(a+bx_i+cx_i^2\right)\right](-1)=0 \quad \ldots.(3)$$

$$\sum_{i=1}^{n} 2\left[y_i-\left(a+bx_i+cx_i^2\right)\right](-x_i)=0 \quad \ldots.(4)$$

and
$$\sum_{i=1}^{n} 2\left[y_i-\left(a+bx_i+cx_i^2\right)\right]\left(-x_i^2\right)=0 \quad \ldots.(5)$$

On further simplification, we get

$$\Sigma y = na + b\Sigma x + c\Sigma x^2 \quad \ldots.(6)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \ldots.(7)$$

and
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \quad \ldots.(7)$$

equations (5), (6) and (7) are called the **"normal equations of the parabola"** solving these equations for a, b and c, we get the best fit for the given equations.

Example 9. Fit a second degree parabola to the following data taking x as the indendent variable.

x:	1	2	3	4	5	6	7	8	9
y:	2	6	7	8	10	11	11	10	9

Solution : Let the equation of parabola for best fit be

$$y = a + bx + cx^2 \quad \ldots.(1)$$

The normal equations are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2 \quad \ldots.(2)$$

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

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \quad \ldots.(4)$$

Here $n = 9$

Table is as follows :

x	y	xy	x^2	x^2y	x^3	x^4
1	2	2	1	2	1	1
2	6	12	4	24	8	16
3	7	21	9	63	27	81
4	8	32	16	128	64	256
5	10	50	25	250	125	625
6	11	66	36	396	216	1296
7	11	77	49	539	343	2401
8	10	80	64	640	512	4096
9	9	81	81	729	729	6561
$\Sigma x = 45$	$\Sigma y = 74$	$\Sigma xy = 421$	$\Sigma x^2 = 284$	$\Sigma x^2 y = 2771$	$\Sigma x^3 = 2025$	$\Sigma x^4 = 15333$

Substituting the value of $\Sigma x, \Sigma y, \Sigma xy, \Sigma x^2, \Sigma x^2 y, \Sigma x^3$ and Σx^4 in (2), (3) and (4), and solving the equations for a, b and c, we get

$$a = -0.923; b = 3.520; c = -0.267$$

Hence the fitted equation is

$$y = -0.923 + 3.520x - 0.267x^2 \quad \textbf{Ans.}$$

Example 10. Using the method of least squares, fit a straight line to the following data

x:	1	2	3	4	5
y:	2	4	6	8	10

Sol. Let the straight line of best fit be

$$y = a + bx \qquad \ldots(1)$$

The normal equations are

$$\left.\begin{aligned} \sum y &= 5a + b\sum x \\ \text{and} \quad \sum xy &= a\sum x + b\sum x^2 \end{aligned}\right\} \qquad \ldots.(2)$$

Table is as below :

x	y	x^2	xy
1	2	1	2
2	4	4	8
3	6	9	18
4	8	16	32
5	10	25	50
$\Sigma x = 15$	$\Sigma y = 30$	$\Sigma x^2 = 55$	$\Sigma xy = 110$

equation (2) becomes

$$30 = 5a + 15b$$

and $$110 = 15a + 55b$$

after solving the above, we get

$$a = 0, b = 2$$

Hence we get the line of best fit is

$$y = 2x$$

Example 11. Fit a parabola $y = a + bx + cx^2$ in least square sense to the data

x:	10	12	15	23	20
y:	14	17	23	25	21

Sol. The normal equations to the curve are

$$\sum y = na + b\sum x + c\sum x^2 \quad \text{...(1)}$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \quad \text{.....(2)}$$

and $$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \quad \text{....(3)}$$

Here $n = 5$

Table is as below :

x	y	x^2	x^3	x^4	xy	x^2y
10	14	100	1000	10000	140	1400
12	17	144	1728	20736	204	2448
15	23	225	3375	50625	345	5175
23	25	529	12167	279841	575	13225
20	21	400	8000	160000	420	8400
$\Sigma x = 80$	$\Sigma y = 100$	$\Sigma x^2 = 1398$	$\Sigma x^3 = 26270$	$\Sigma x^4 = 521202$	$\Sigma xy = 1684$	$\Sigma x^2 y = 30648$

Now, substituting all the values in normal equations, we get

$$100 = 5a + 80b + 1398c$$

$$1684 = 80a + 1398b + 26270c$$

and $30648 = 1398a + 26270b + 521202c$

after solving the above equations, we get

$$a = -8.89, b = 3.03 \text{ and } c = -0.07$$

Hence the required equation is

$$y = -8.8a + 3.03b - 0.07x^2 \quad \textbf{Ans.}$$

Example 12. Fit a second degree parabola to the following data :

x	1	2	3	4	5
y	1090	1220	1390	1625	1915

Sol. Let us define u and v s.t.

$$u = x - 3 \text{ and } v = \frac{y - 1450}{5}$$

and equation of the parabola

$$v = a + bu + cu^2 \qquad \ldots.(1)$$

x	y	$u=(x-3)$	$v=\frac{y-1450}{5}$	u^2	u^4	uv	u^2v
1	1090	–2	–72	4	16	144	–288
2	1220	–1	–46	1	1	46	–46
3	1390	0	–12	0	0	00	00
4	1625	1	35	1	1	35	35
5	1915	2	93	4	16	186	372
		$\Sigma u = 0$	$\Sigma V = -2$	$\Sigma u^2 = 10$	$\Sigma u^2 = 34$	$\Sigma uv = 411$	$\Sigma u^2 v = 73$

The normal equations are

$$\sum v = na + b\sum u + c\sum u^2 \qquad \ldots.(2)$$

$$\sum uv = a\sum u + b\sum u^2 + c\sum u^3 \qquad \ldots.(3)$$

and $$\sum u^2 v = a\sum u^2 + b\sum u^3 + c\sum u^4 \qquad \ldots.(4)$$

where $n = 5$

Substituting all the values from the above table in normal equations, we get

$$-2 = 5a + b(0) + 10c$$

$$411 = a(0) + 10b + c(0)$$

and $$73 = 10a + b(0) + c(34)$$

On solving, we obtain

$$a = -11.4, b = 41.1 \text{ and } c = 5.5$$

$\therefore$ (1) become

$$V = -11.4 + 41.1(u) + 5.5(u^2)$$

or $$\left(\frac{y-1450}{5}\right) = -11.4 + 41.1(x-3) + 5.5(x-3)$$

$\Rightarrow$ $$y = 1024 + 40.5x + 27.5x^2$$ **Ans.**

11.11 REGRESSION ANALYSIS

11.11.1 Introduction

The term "regression" literally means "stepping back towards the average". It was first used by **British biometrician Sir Francis Galton (1822-1911)**. In connection with the inheritance of stature francis found that the offsprings of abnormally tall or short parents tend to **"regress"** or **"step back"** to the average population height, but the term **"regression"** as now, used in statistics in only a convenient term without having any reference to biometry.

Def. Regression analystis is a mathematical measure of the average relationship between two or more variables in terms of the original units of data.

There are two types of the variables use in regression analysis. The variable whose value is to be predicted is called dependent variable and other is used for prediction is called independent variable.

In regression analysis independent variable is called regressor or predictor or explanatory variable while the dependent variable is known as regressed or expalained variable.

11.11.2 Line of Regression

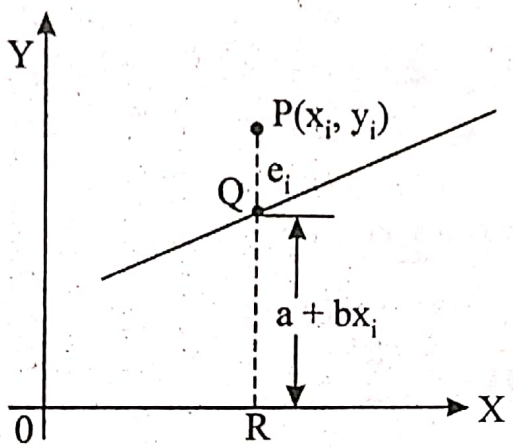
If the scatter diagram indicates some relations between two variables x and y, then the dots of the scatter diagram will be concentrated round a curve. This cure is called the **curve of regression**. When the curve is straight line, it is called a **line of regression**. A line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

11.11.3 Equations of the Lines of Regression

Suppose $y = a + bx$ ...(1)

be the equation of the **line of regression of y on x.**

Let (x_i, y_i) be any point of dot according to the given figure



$$PR = y_i$$

$$QR = a + bx_i$$

$$\therefore \quad PQ = PR - QR$$

$$e_i = y_i - (a + bx_i)$$

is called the residual for ith point or error of estimate for y_i.

Let S be the sum of the square of such distances.

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (a + bx_i) \right]^2$$

According to the principle of least squares, we have to choose a and b s.t. S is minimum.

The method of least square gives the condition for minimum value of S.

$$\frac{\partial S}{\partial a} = 2\sum_{i=1}^{n} \left[\left(y_i - (a + bx_i) \right) \right](-1) = 0$$

$$\Rightarrow \quad \sum y = na + b\sum x \qquad \text{....(2)}$$

and $$\frac{\partial S}{\partial b} = 2\sum_{i=1}^{n} \left[y_i - (a + bx_i) \right](-x_i)$$

$$\Rightarrow \quad \sum xy = a\sum x + b\sum x^2 \qquad \text{....(3)}$$

equation (2) and (3) are called normal equations now dividing (2) by n we get

$$\frac{\Sigma y}{n} = a + b\frac{\Sigma x}{n} \qquad \left(\because \bar{y} = \frac{\Sigma y}{n}, \bar{x} = \frac{\Sigma x}{n}\right)$$

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

This shows that $(\bar{x}, \bar{y})$ lie on the line of regression (1), shifting the origin to $(\bar{x}, \bar{y})$, the equation (3) becomes

$$\sum(x-\bar{x})(y-\bar{y}) = a\sum(x-\bar{x}) + b\sum(x-\bar{x})^2$$

But $\sum(x-\bar{x}) = 0$

i.e. $$\sum(x-\bar{x})(y-\bar{y}) = b\sum(x-\bar{x})^2$$

or $$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum XY}{\sum X^2} \qquad \ldots.(5)$$

$\because$ we know that

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2}\sqrt{\Sigma Y^2}} = \frac{\Sigma XY}{n\sqrt{\frac{\Sigma X^2}{n}}\sqrt{\frac{\Sigma Y^2}{n}}}$$ [Here $\frac{\Sigma XY}{n}$ = covariance (x, y) σ_x and σ_y are standard deviation]

or $$\Sigma XY = n.r.\,\sigma_x.\sigma_y$$

Putting the value of $\sum XY$ in (4) we get

$$b = \frac{nr\sigma_x\sigma_y}{\Sigma X^2} = \frac{n\sigma_x\sigma_y}{\frac{\Sigma X^2}{n}} = \frac{n\sigma_x\sigma_y}{\sigma_x^2} = \frac{r\sigma_y}{\sigma_x}$$

Therefore the slope of the line of regression

$$b = \frac{r\sigma_y}{\sigma_x}$$

Hence the equation of the **line of regression y on x**, passes through $(\bar{x}, \bar{y})$ is

$$\boxed{y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})} \qquad \ldots.(6)$$

Similarly **the line of regression x on y** is

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \ldots(7)$$

Note:

(1) If $r = 0$ in equation (6) and (7), the line of regression becomes $x = \bar{x}$ and $y = \bar{y}$ which are two straight lines parallel to y and x axis respectively and passing through $\bar{x}$ and $\bar{y}$. They are mutually perpendicular and if $r = \pm 1$, the two lines of regression will coincide.

(2) The regression coefficient X on Y is represented by b_{xy} and the regression coefficient Y on X is represent by b_{yx} i.e.

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

and $$b_{xy} \cdot b_{yx} = \left(\frac{r\sigma_x}{\sigma_y}\right) \cdot \left(\frac{r\sigma_y}{\sigma_x}\right) = r^2$$

11.11.4 Use of Regression Analysis

1. Regression analysis is used in the field of business, this tool of statistical analysis is widely used. Businessmen are interested in predicting future production, consupmption, investment, prices, profits and sales etc.
2. In the field of economic planning and sociological studies, projections of pupulation, birth rates, death rates and other similar variables are of great use.

Example 13. If θ be the acute angle between the two regression lines in the case of two variables x and y, show that

$$\tan\theta = \frac{1 - r^2}{r} \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

Where r, σ_x, σ_y have their usual meanings. Explain the significance where r = 0 and $r = \pm 1$. **[Raj. Univ. BE (CS) 2003]**

Solution: Lines of regression are

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \ldots(1)$$

$$\therefore \quad m_1 = r\frac{\sigma_y}{\sigma_x}$$

and $x-\bar{x}=\frac{r\sigma_x}{\sigma_y}(y-\bar{y})$ ...(2)

$\therefore \quad m_2=\frac{1}{r}\frac{\sigma_y}{\sigma_x}$

$$\tan\theta=\frac{m_2-m_1}{1+m_1m_2}$$

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

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

$$\Rightarrow \quad \tan\theta=\frac{1-r^2}{r}\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \qquad ....(3)$$

(i) If r = 0, then there is no relationship between the two variables and they are independent.

On putting r = 0 in equation (3), we get $\tan\theta=\infty, \theta=\frac{\pi}{2}$, so the line (1) and (2) are perpendicular.

(ii) If $r=\pm 1$

on putting these values of r in (3), we get

$\tan\theta=0$ or $\theta=0$

i.e. lines (1) and (2) are concide. The correlation is perfect. **Ans.**

Example 14 If the regression coefficient are 0.8 and 0.2, what would be the value of coefficient of correlation ?

Solution : Here $b_{yx}=0.8$ and $b_{xy}=0.2$

But we know that

$$r^2=b_{yx}\cdot b_{xy}=0.8\times 0.2=0.16$$

Hence $r=0.4$

Example. 15 From the following data, find the most likely value of y when x = 24 :

	y	x
Mean	985.8	18.1
S.D.	36.4	2.0

and $r = 0.58$

Sol. $\because \bar{x} = 18.1, \bar{y} = 985.8, \sigma_y = 36.4, \sigma_x = 2.0$ and $r = 24$ 0.58

$\therefore$ Regression Coefficient $\sigma_{yx} = r\dfrac{\sigma_y}{\sigma_x}$

$$= (24)\frac{36.4}{2.0} = 10.556.$$

Regression line y on x is

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

$$y - 985.8 = (10.556)(x - 18.1)$$

$$\Rightarrow \quad y = 10.556x + 794.73$$

Now when x = 24

$$y = 10.556 \times 24 + 794.73$$

$$y = 1048$$

Example. 16 The equations of two regression lines, obtains in a correlation analysis of 60 observations are :

$$5x = 6y + 24 \text{ and } 1000y = 786x - 3608$$

What is the correlation coefficient? Show the ratio of coefficient of variability of x to that of y is $\frac{5}{24}$. What is the ratio of variances of x and y?

Sol. Here Regression line x and y is

$$5x = 6y + 24$$

$$\Rightarrow \quad x = \frac{6}{5}y + \frac{24}{5} \qquad \ldots(1)$$

$$\therefore \quad b_{xy} = \frac{6}{5} \qquad \ldots.(2)$$

Regression line y on x is

$$1000y = 768x - 3608$$

$$y = \frac{768}{1000}x - \frac{3608}{1000} \quad \ldots(3)$$

$$\therefore \quad b_{yx} = \frac{768}{1000} \quad \ldots(4)$$

From (2) $b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{6}{5}$ $\ldots(5)$

and from (4) $b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.768$ $\ldots(6)$

Mulitplying (5) and (6), we get

$$r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = \frac{6}{5} \times 0.768$$

$$r^2 = 0.9216 \Rightarrow r = 0.96 \quad \ldots(7)$$

Now divide (6) by (5)

$$\frac{\sigma_x^2}{\sigma_y^2} = \frac{6}{5 \times 0.768} = 1.5625$$

$$\Rightarrow \quad \frac{\sigma_x}{\sigma_y} = \sqrt{1.5625} = 1.25 = \frac{5}{4}$$

$\because$ We know that the regression line passes through the point $(\bar{x}, \bar{y})$, we have

$$5\bar{x} = 6\bar{y} + 24$$

$$1000\bar{y} = 768\bar{x} - 3608$$

Solving the above equations, we get

$$\bar{x} = 6 \text{ and } \bar{y} = 1$$

$\therefore$ Coefficient of variability of $x = \frac{\sigma_x}{\bar{x}}$

and Coefficient of variability of $y = \frac{\sigma_y}{\bar{y}}$

Hence required ratio $= \frac{\sigma_x}{\bar{x}} \cdot \frac{\bar{y}}{\sigma_y}$

$$\frac{\bar{y}}{\bar{x}} \cdot \frac{\sigma_x}{\sigma_y} = \frac{1}{6} \times \frac{5}{4} = \frac{5}{24}$$

Example 17. In a corrlelation study, the followign values are obtained :

	x	y
Mean	65	67
S.D.	2.5	3.5

Coefficient of correlation r = 0.8

Find the two regression equations

Sol. ∵ Regression equation x and y.

$$x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

Here $\bar{x}=65,\sigma_x=2.5,\sigma_y=3.5,r=0.8$ and $\bar{y}=67$.

$$\therefore \quad x-65=0.8\left(\frac{2.5}{3.5}\right)(y-67)$$

$$x-65=0.571y-38.26$$

$\Rightarrow$ $x=0.571y+26.74$, which is required result.

We also know that regression equation y on x is

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$y-67=(0.8)\left(\frac{3.5}{2.5}\right)(x-65)$$

$$y-67=1.12x-72.8$$

$$y=1.12x-5.8$$

which is required result.

Example 18. Find the two lines of regression and coefficient of correlation for the data given below:

$n=18,\sum x=12,\sum y=18,\sum x^2=60,\sum y^2=96,\sum xy=48$ **[Raj. Univ. CS 2006]**

Solution : Here $n = 18$

and $\bar{x}=\dfrac{\sum x}{n}=\dfrac{12}{18}=0.667$

$$\bar{y} = \frac{\sum y}{n} = \frac{18}{18} = 1$$

Coefficient of correlation

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}\ \sqrt{\frac{1}{n}\sum y_i^2 - \bar{y}^2}}$$

$$= \frac{\frac{48}{18} - 0.667 \times 1}{\sqrt{\frac{60}{18} - (0.667)^2}\ \sqrt{\frac{96}{18} - 1}}$$

$$= \frac{2.66 - 0.667}{\sqrt{3.33 - 0.444}\ \sqrt{5.33 - 1}} = \frac{1.993}{6.005} = 0.331$$

$$r = 0.331$$

Here $\sigma_x = 2.886$, $\sigma_y = 2.081$

Regression equation of y on x

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$(y - \bar{y}) = 0.33\left(\frac{2.081}{2.886}\right)(x - 0.667)$$

$$(y - 1) = 0.237\,(x - 0.667)$$

$$\boxed{y = 0.237x + 0.9436}$$ **Ans.**

Regression equation of x on y

$$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$(x - 0.667) = 0.331\left(\frac{2.886}{2.081}\right)(y - 1)$$

$$x - 0.667 = 0.4590\,(y - 1)$$

$$\boxed{x = 0.459y + 0.2079}$$ **Ans.**

Example 19. Calculate the coefficient of correlation and obtain the lines of regression for the following data.

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

Solution:

x	y	x^2	y^2	xy
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
$\Sigma x = 45$	$\Sigma y = 108$	$\Sigma x^2 = 285$	$\Sigma y^2 = 1356$	$\Sigma xy = 597$

Here $n = 9$, $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{45}{9} = 5$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{108}{9} = 12$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum_i x_i^2 - \bar{x}^2} = \sqrt{\frac{285}{9} - 25} = \sqrt{\frac{60}{9}}$$

$$\sigma_x = 2.58$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_i y_i^2 - \bar{y}^2} = \sqrt{\frac{1356}{9} - 144} = \sqrt{\frac{60}{9}}$$

$$\sigma_y = 2.58$$

$$r = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}\ \sqrt{\frac{1}{n}\sum y_i^2 - \bar{y}^2}}$$

$$= \frac{\frac{597}{5} - 60}{\sqrt{\frac{60}{9}}\sqrt{\frac{60}{9}}}$$

$\boxed{r = 0.95}$ **Ans.**

The regression line y on x then

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$(y - 12) = 0.95\left(\frac{2.58}{2.58}\right)(x - 5)$$

$$(y - 12) = 0.95x - 4.75$$

$\boxed{y = 0.95x + 7.25}$ **Ans.**

and line of regression x on y

$$(x - \bar{x}) = r\frac{\sigma_y}{\sigma_y}(y - \bar{y})$$

$$(x - 5) = 0.95\left(\frac{2.58}{2.58}\right)(y - 12)$$

$\boxed{x = 0.95y - 6.4}$ **Ans.**

Example 20. In a partially destroyed laboratory on record of an analysis of correlation data, the following results only are legible, variance $(x) = 9$, regression equations are $8x - 10y + 66 = 0$ and $40x - 18y - 214 = 0$ find

(i) The mean values of x and y.

(ii) The coefficient of correlation between x and y.

(iii) The standard deviation of y.

Solution : (i) The mean value is the common point of intersection of the two lines of regression

$$8x - 10y = -66 \quad \ldots(1)$$

and $40x - 18y = 214$ ....(2)

Solving the equation (1) and (2), we get the mean values of x and y. i.e.

$x = \bar{x} = 13$ and $y = \bar{y} = 17$.

(ii) The line of regression can be written as

$$y = \frac{8}{10}x + \frac{66}{10} \quad \text{(regression } y \text{ on } x\text{)}$$

and $x = \frac{18}{40}y + \frac{214}{40}$ (regression x on y)

Here $b_{xy} = \frac{18}{40}$ and $b_{yx} = \frac{8}{10}$

$\therefore \quad r^2 = b_{xy} \times b_{yx} = \frac{18}{40} \times \frac{8}{10} = \frac{9}{25} = 0.36$

$\therefore \quad r = \pm 0.6$

$\because$ b_{xy} and b_{yx} both the regression coefficients are positive therefore correlation coefficient r should be positive i.e. r = 0.6

(iii) Again $b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{8}{10}$

Variance $(x) = 9$ (given)

$\therefore \quad S.D. = \sqrt{\text{var}\,iance(x)} = \sqrt{9} = 3$

i.e. $\sigma_x = 3$

$\Rightarrow \quad \frac{8}{10} = 0.6\frac{\sigma_y}{3}$

$\therefore \quad \sigma_y = \frac{8 \times 3}{10 \times 0.6}$

$\sigma_y = 4$

Example 21. Two random variables have the least square regression lines with equations $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$. Find the mean values and the coefficient of correlation between x and y. **[Raj. Univ. CS 2003, 2006]**

Solution: Given lines of regression

$3x + 2y = 26 \quad ...(1)$

$6x + y = 31 \quad ...(2)$

Solving (1) and (2), we get $x = 4, y = 7$

Hence the mean values are $\bar{x} = 4, \bar{y} = 7$

We know that the lines of regression

$$(y-\bar{y}) = r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \quad \ldots(3)$$

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \quad \ldots(4)$$

we can write equation (1) and (2)

$$y = -\frac{3}{2}x+13 \quad \ldots(5)$$

$$x = -\frac{y}{6}+\frac{31}{6} \quad \ldots(6)$$

Comparison of (3) and (5), (4) and (6)

$$r\frac{\sigma_y}{\sigma_x} = -\frac{3}{2} = b_{yx}$$

$$r\frac{\sigma_x}{\sigma_y} = -\frac{1}{6} = b_{xy}$$

Correlation of coefficient

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

$$= -\frac{3}{2}\times\frac{-1}{6} = \frac{1}{4}$$

$$r = \pm\frac{1}{2}$$

$\boxed{r = \pm 0.5}$ **Ans.**

Example 22. Calculate the coefficient of rank correlation from the following data

x	48	33	40	9	16	16	65	24	16	57
y	13	13	24	6	15	4	20	9	6	19

Solution:

x	y	Rank of x d_1	Rank of y d_2	$d = d_1 - d_2$	d^2
48	13	3	5.5	–2.5	6.25
33	13	5	5.5	–0.5	0.25
40	24	4	1	3	9
9	6	10	8.5	1.5	2.25
16	15	8	4	4	16
16	4	8	10	–2	4
65	20	1	2	–1	1
24	9	6	7	–1	1
16	6	8	8.5	–.5	25
57	19	2	3	–1	1
				$\sum d = 0$	$\sum d^2 = 41$

Since ranks are repeated then rank correlation coefficient

$$r = 1 - \frac{6\left\{\sum d_i^2 + \frac{m\left(m^2-1\right)}{12}\right\}}{n\left(n^2-1\right)} = 1 - \frac{6\left\{\sum d_i^2 + F\right\}}{n\left(n^2-1\right)}$$

First we calculated F for repeated rank (separately for x and y) in x-term

16 repeated thrice $(m_1 = 3)$

then $$F_1 = \frac{m_1\left(m_1^2-1\right)}{12} = \frac{3(9-1)}{12} = \frac{3\times 8}{12} \qquad [\because n = 10]$$

$$F_1 = 2$$

In y-term 13 repeated twice $(m_2 = 2)$

6 repeated twitce $(m_3 = 2)$

$$F_2 = \frac{m_2\left(m_2^2-1\right)}{12} + \frac{m_3\left(m_3^2-1\right)}{12}$$

$$= \frac{2(4-1)}{12} + \frac{2(4-1)}{12} = 1$$

then $F = F_1 + F_2$

$$= 2 + 1 = 3$$

so, $$r = 1 - \frac{6\{41+3\}}{10(100-1)}$$

$$= 1 - \frac{6 \times 44}{10 \times 99} = 1 - \frac{264}{990}$$

$\boxed{r = 0.7334}$ **Ans.**

Example 23. Calculate the linear regression coefficients from the following data.

x	1	2	3	4	5	6	7	8
y	3	7	10	12	14	17	20	24

Solution:

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 $n = 8$,

Since we know that

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_{yx} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} \times \frac{\sigma_y}{\sigma_x}$$

$$b_{yx} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sigma_x^2} \qquad \ldots(1)$$

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

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} \times \frac{\sigma_x}{\sigma_y}$$

$$b_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sigma_y^2} \qquad \ldots.(2)$$

Here $n = 8$, $\bar{x} = \frac{1}{n}\sum x_i = \frac{36}{8} = 4.5$ $\quad \bar{x} = \frac{\Sigma x}{n}$

$\bar{y} = \frac{1}{n}\sum y_i = \frac{107}{8} = 13.37$ $\quad \bar{y} = \frac{\Sigma y}{n}$

$$\sigma_x^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

$$= \frac{204}{8} - (4.5)^2 = 25.5 - 20.25$$

$$\sigma_x^2 = 5.25$$

$$\sigma_y^2 = \frac{1}{n}\sum y_i^2 - \bar{y}^2$$

$$= \frac{1763}{8} - (13.37)^2 = 220.37 - 178.75$$

$$\sigma_y^2 = 41.62$$

from (1) $$b_{yx} = \frac{\frac{599}{8} - 4.5 \times 13.37}{41.62} = 74.875 - 60.165$$

$$= \frac{14.71}{41.62} = 0.353$$

$\boxed{b_{yx} = 0.353}$ **Ans.**

$$b_{xy} = \frac{14.71}{5.25}$$

$\boxed{b_{xy} = 2.80}$ **Ans.**

Example 24. A panel of two judges, A and B graded seven T.V. serial performances by awarding marks independently as shown in the following table.

Performance	1	2	3	4	5	6	7
Marks by A (x)	46	42	44	40	43	41	45
Marks by B (y)	40	38	36	35	39	37	41

The eight T.V. performance which judge B could not attend, was awarded 37 marks by judge A. If the judge B has also been present, how many marks would be expected to have been awarded by him to the eight T.V. performance.Use regression analysis to answer the above question.

Solution: Let the marks awarded by judge A be denoted by x and marks awarded by judge B be denoted by y.

x	y	x^2	xy
46	40	2116	1840
42	38	1764	1596
44	36	1936	1584
40	35	1600	1400
43	39	1849	1677
41	37	1681	1517
45	41	2025	1845
$\Sigma x = 301$	$\Sigma y = 266$	$\Sigma x^2 = 12971$	$\Sigma xy = 11459$

Here $n = 7$, $\bar{x} = \frac{\sum x}{n} = \frac{301}{7} = 43$

$\bar{y} = \frac{\sum y}{n} = \frac{266}{7} = 38$

Regression coefficient

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{1}{n}\frac{\sum x_i y_i - \bar{x}\,\bar{y}}{\sigma_x^2}$$

Since $$\sigma_x^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

$$= \frac{12971}{7} - (43)^2 = \frac{12971 - 12943}{7}$$

$$\sigma_x^2 = 4$$

then $$b_{yx} = \frac{\frac{11459}{7} - 43\times 38}{4} = \frac{11459 - 11438}{28}$$

$$= \frac{3}{4} = 0.75$$

$$\boxed{b_{yx} = 0.75}$$ **Ans.**

Regression line y on x then

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

$$y - 38 = 0.75(x - 43)$$

$$y = 0.75x + 5.75$$

if $$x = 37$$

$$y = .75 \times 37 + 5.75$$

$$y = \textbf{33.5 marks}$$ **Ans.**

Hence if judge B had also been presents 33.5 marks would be expected to have been awarded to the eight T.V performance.

Example 25. Calculate the coefficient of correlation between x and y using the following data:

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

Solution :

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
1	9	–4	–3	16	9	12
2	8	–3	–4	9	16	12
3	10	–2	–2	4	04	04
4	12	–1	0	1	00	–01
5	11	0	–1	0	01	–01
6	13	1	1	1	01	01
7	14	2	2	4	04	04
8	16	3	4	9	16	12
9	15	4	3	16	09	12
$\Sigma x = 45$	$\Sigma y = 108$			$\Sigma X^2 = 60$	$\Sigma Y^2 = 60$	$\Sigma XY = 57$

$$\because \quad \bar{x} = \frac{\sum x}{n} = \frac{45}{a} = 5; \qquad \bar{y} = \frac{\sum y}{n} = \frac{108}{9} = 12$$

Karl pearson coefficient of correlation

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} = \frac{57}{\sqrt{60 \times 60}} = \frac{57}{60}$$

= 0.95 Ans.

Example 26: The ranks of same 16 students in mathematics and physics are as follows. Two numbers within brackets denote the ranks of the students in mathematics and physics: (1, 1), (2, 10), (3, 3), (4, 4), (5, 5), (6, 7), (7, 2), (8, 6), (9, 8), (10, 11), (11, 15), (12, 9), (13, 14), (14, 12), (15, 16), (16, 13).

Calculate the rank correlation coefficient for proficiencies of this group in mathematics and physics.

Solution :

Ranks in Maths (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	**Total**
Ranks in Physics (Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	15	16	
$d = X - Y$	0	–8	0	0	0	–1	5	2	1	–1	–4	3	–1	2	–1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Rank correlation coefficient is given by

$$\rho = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6 \times 136}{16 \times 255} = 1 - \frac{1}{5}$$

$$= \frac{4}{5} = 0.8 \quad \textbf{Ans.}$$

EXERCISE – 11.1

Q.1 Calculate the coefficient of correlation between x and y using the following data.

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

Ans. $r_{xy} = 0.95$

Q.2 Calculate the Pearson's coefficient of correlation between x and y.

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Ans. $r_{xy} = 0.97$

Q.3 Calculate the coefficient of correlation between the values of x and y given following data.

x	78	89	79	69	59	79	68	61
y	125	137	156	112	107	136	123	108

Ans. $r_{xy} = 0.95$

Q.4 Show that the following data are uncorrelated.

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

Ans. $r_{xy} = 0$

Q.5 Calculate Karl Pearson's correlation coefficient between the values of x and y, given following data.

x	2	5	7	9	19	17
y	25	27	26	29	34	35

Ans. $r_{xy} = 0.976$

Q.6 Calculate the coefficient of correlation for the following ages of husband and wife.

Husband's age	23	27	28	29	30	31	33	35	36	39
Wife's age	18	22	23	24	25	26	28	29	30	32

Ans. $r = 0.95$

Q.7 A sample of 12 father's and their eldest sons gave the following data about their height in inches.

Father	65	63	67	64	68	62	70	66	68	67	69	71
Son	68	66	68	65	69	66	68	65	71	67	68	70

Calculate the ranks correlation coefficient.

Ans. *0.72*

Q.8 The following are the number of hours which 10 students studied for an examination and scores they obtained.

No. of hours studied	8	5	11	13	10	5	18	14	3	8
Scores	56	44	79	72	70	54	94	86	32	65

Calculate the rank correlation coefficient.

Ans. *0.92*

Q.9 Find a regression lines of x on y for the following data.

x	1	5	3	2	1	1	7	3
y	6	1	0	0	1	2	1	5

Ans. $72x = -20y + 247$

Q.10 Two lines regression are given by

$x + 2y = 5$, and $2x + 3y = 8$ and variance of $x = 12$. Find

(i) The mean values of x and y

(ii) Variance of y

(iii) The coefficient of correlation between x and y.

Ans. *(i)* $\bar{x} = 1, \bar{y} = 2$ *(ii) 4 (iii)* $\frac{-\sqrt{3}}{2}$

Q.11 Find the correlation coefficient between x and y for the given values if $n = 15$, $\Sigma x = 50$, $\Sigma y = -30$, $\Sigma x^2 = 290$, $\Sigma y^2 = 300$, $\Sigma xy = -115$.

Q.12 Find the regression lines by using the following data

$\bar{x} = 0$, $\bar{y} = 22$, $\sigma_x = 4$, $\sigma_y = 5$, $r = 0.9$

Estimate y when $x = 25$.

Ans. *50.125*

Q.13 Find correlation coefficient r when it is given that the two regression coefficients are 0.8 and 1.2.

Ans. *0.92*

Q.14 Find the regression lines and coefficient of correlation between x and y from the following data. **[Raj. CP 2002]**

x	45	55	56	58	60	65	68	70	75	80	85
y	56	50	48	60	62	64	65	70	74	82	90

Ans. *0.92, $x = 0.85y + 9.47$, $y = .99x + 1.02$*

Q.15 The following gives the data of rainfall and discharge in a certain river obtain the line of regression of y on x.

Rain fall	1.5	1.8	2.6	2.9	3.4
Discharge	33	36	40	46	53

Ans. $y = 17.35 + 9.94x$

Q.16 Two random variables have the least square regression lines with equations $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$. Find the mean values and the coefficient of correlation between x and y.

Ans. $\bar{x} = 4, \bar{y} = 7, r = -0.5$

Q.17 Obtain regression line of x on y for the given data.

x	1	2	3	4	5	6
y	5.0	8.1	10.6	13.1	16.2	20.0

Ans. $x = 0.341\ y - 0.66$