Advanced Engineering Mathematics (CS-Branch-III Sem)

Prof.(Dr.) Ashok Singh Shekhawat
Department of Mathematics
JECRC, Jaipur

RAJASTHAN TECHNICAL UNIVERSITY, KOTA

Syllabus

II Year-III Semester: B.Tech. Computer Science and Engineering

3CS2-01: Advanced Engineering Mathematics

Credit-3
3L+0T+0P

Max. Marks : 150 (IA:30,ETE:120)
End Term Exam: 3 Hours

SN	CONTENTS	Hours
1	**Random Variables:** Discrete and Continuous random variables, Joint distribution, Probability distribution function, conditional distribution.Mathematical Expectations: Moments, Moment Generating Functions, variance and correlation coefficients, Chebyshev's Inequality, Skewness and Kurtosis.	7
2	**Binomial distribution**, Normal Distribution, Poisson Distribution and their relations, Uniform Distribution, Exponential Distribution. Correlation: Karl Pearson's coefficient, Rank correlation. Curve fitting. Line of Regression.	5
3	**Historical development**, Engineering Applications of Optimization, Formulation of Design Problems as a Mathematical Programming Problems, Classification of Optimization Problems	8
4	**Classical Optimization using Differential Calculus:** Single Variable and Multivariable Optimization with & without Constraints, Langrangian theory, Kuhn Tucker conditions	6
5	**Linear Programming:** Simplex method, Two Phase Method and Duality in Linear Programming. Application of Linear Programming: Transportation and Assignment Problems.	14
	TOTAL	40

unit - II

Probability Distribution

① Binomial Distribution
② Poissian Distribution
③ Normal Distribution.

Binomial Distribution

suppose a random experiment is performed repeatedly. let E be an event, we shall call the occurrence of the event E, a success and its non-occurrence a failure. If p denotes the Probabity of a success and q denotes the Probability of its failure, then $p+q=1$ let the event E be Tried n times, where n is finite, The hypotheses for Binomial distribution are

(i) All the Trials are Independent, i.e the Result of one Trial will not affect the results of succeeding trials.

(ii) The number (n) of Trials is finite.

(iii) The Probability p of successes is the same in every Trial.

Now the Probability of r Trials outs of n Trials. is given by the formula. of Binomial distribution.

$$\boxed{P(r) = n_{C_r}\, p^r q^{n-r}} \qquad r = 0, 1, 2, 3 - - -$$

$$\boxed{p+q=1} \qquad \boxed{q=1-p}$$

Since we have considered N sets, each of n trials, Therefore the number of sets with r successes $= N \cdot {}^{n}C_{r}\, p^{r} q^{n-r}$ or $N \cdot (P(r))$

$r = 0, 1, 2\ 3 - - - -$

** Mean, Variance, and Standard deviation of Binomial Distribution

(i) First Moment about the origin :-

$$\mu_1' = \sum_{r=0}^{n} r \cdot {}^{n}C_{r}\, p^{r} q^{n-r} \quad \text{where } p+q=1$$

$$= \sum_{r=1}^{n} n\ {}^{n-1}C_{r-1}\, p^{r} q^{n-r}$$

$$\left[\because r\, {}^{n}C_{r} = n \cdot {}^{n-1}C_{r-1}\right]$$

$$= \sum_{r=1}^{n} np\ {}^{n-1}C_{r-1}\, p^{r-1} q^{(n-1)-(r-1)} = np \text{ (Binomial Distribution)}$$

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

$$\boxed{\mu_1' = np}$$

$$\boxed{\text{Mean} = \mu_1' = np}$$

(ii) Variance :- (σ^2) second Moment about origin :-

$$\mu_2' = \sum_{r=0}^{n} r^2 p^{r} q^{n-r}\, {}^{n}C_{r}$$

$$\mu_2' = \sum_{r=0}^{n}[r(r-1)+r]\, {}^{n}C_r\, p^r q^{n-r}$$

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

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

$$\because \left[r(r-1)\, {}^{n}C_r = n(n-1)\, {}^{n-2}C_{r-2}\right]$$

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

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

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

$n^2p^2 - np^2 + np$

$\mu_2' = n^2p^2 + np(1-p) = n^2p^2 + npq$

μ_2' = second Moment about origin.

$\mu_2' = n^2p^2 + npq$

variance $\sigma^2 = \mu_2' - \mu_1'^2 = \mu_2$

$\sigma^2 = \cancel{n^2p^2} + npq - \cancel{n^2p^2}$

$\boxed{\sigma^2 = npq}$

$S.D = \sqrt{\text{Variance}} = \sqrt{npq}$

II - 01/32

constitute a discrete probability distribution for X and it spells out how the total probability of 1 is distributed over several values of the random variable.

4.3 MATHEMATICAL EXPECTATION

If X is a discrete random variable which takes the possible values $x_1, x_2, x_3 \ldots\ldots x_n$ with respective probabilities $p_1, p_2, p_3 \ldots\ldots p_n$ (where each $p_i \geq 0$ and $\Sigma p_i = 1$) then its mathematical expectation or simply the expectation is defined as

$$E(X) = p_1 x_1 + p_2 x_2 + p_n x_n = \sum_{i=1}^{n} p_i x_i.$$

If $\phi(X)$ is some function of the variate X such that it takes values $\phi(x_1), \phi(x_2), \ldots, \phi(x_n)$ when X takes the values $x_1, x_2, x_3 \ldots\ldots x_n$ then its expectation is given by

$$E[\phi(X)] = p_1\phi(x_1) + p_2\phi(x_2) + \ldots + p_n\phi(x_n) \qquad \ldots(1)$$

$$= \sum_{i=1}^{n} p_i\phi(x_i) \quad \text{where } \Sigma p_i = 1$$

(i) If $\phi(X) = X^r$ then (1) gives

$$E(X^r) = p_1 x_1^r + p_2 x_2^r + \ldots\ldots p_n x_n^r$$

$$= \sum_{i=1}^{n} p_i x_i^r = \sum_{i=1}^{n} p_i (x_i - 0)^r$$

which is defined as μ_r', the r^{th} moment about origin of the discrete probability distribution.

Thus μ_r', (about origin) $= E(X^r) = \Sigma p_i x_i^r$ ....(2)

In particular, $\mu_1' = E(X) = p_1 x_1 + p_2 x_2 + \ldots + p_n x_n = \sum_{i=1}^{n} p_i x_i$

If p_i is replaced by $\dfrac{f_i}{N}$ where $\sum_{i=1}^{n} f_i = N$, then

$$E(X) = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \text{mean} = \bar{x} = \mu_1'$$

Here $E(X)$ denotes the mean in random experiment when X takes the values $x_1, x_2 \ldots\ldots x_n$ with frequencies $f_1, f_2, \ldots f_n$. This moment $E(X)$ is called the mean of the variate or the distribution.

(ii) If $\phi(X) = [X-E(X)]^r$

then $E[\{X-E(X)\}^r] = E[(X-\overline{x})^r] = \sum_{i=1}^{n} p_i(x_i-\overline{x})^r$

which is μ_r, the r^{th} moment about mean.

$\therefore\ \mu_r = E[(X-\overline{x})^r] = \sum_{i=1}^{n} p_i(x_i-\overline{x})^r$

In particular if $r = 2$, we get $\mu_2 = E[(X-\overline{x})^2] = \sum_{i=1}^{n} p_i(x_i-\overline{x})^2$

which is defined as variance of X or var(X). If $\overline{x}$ is not a whole number, then

$$\mu_2 = \Sigma p_i(x_i-\overline{x})^2 = \Sigma p_i x_i^2 - (\overline{x})^2$$

(iii) If X is a continuous variate with the probability density $f(x)$ then its mathematical expectation is given by

$$E(X) = \int_a^b x\,f(x)dx \qquad \begin{bmatrix}\text{where } a \le X \le b \text{ and} \\ P(x \le X \le x+dx) = f(x)dx\end{bmatrix}$$

In general, the mathematical expectation of $\phi(X)$, the random variable, whose value is $f(x)$ when the value of X is x given by

$$E[\phi(X)] = \int_a^b \phi(x)\,f(x)dx$$

In particular,

$$E[\{X-E(X)\}^r] = \mu_r = \int_a^b (x-\overline{x})^r\ f(x)dx \quad (\text{for } r = 1, 2....)$$

and $E(X^r) = \mu'_r = \int_a^b x^r f(x)dx$ (for $r = 1, 2, 3....$)

Addition theorem of Expectation: If X, Y, Z .... T are n random variates, then $E(X+Y+Z+.....+T) = E(X)+E(Y)+E(Z)+......+E(T)$ if all the expectations on the right exist.

Independent Variate: Two random variables X and Y on a sample space are called independent if the probability that either of them does not dependent on the other.

Multiplication theorem of Expectation: The mathematical expectation of the product of a number of independent random variables to equal is the product of their expectation. Symbolically, if X, Y, Z .... T are n independent random variables, then

$$E(XYZ.....T) = E(X)E(Y)E(Z)....E(T)$$

Covariance: If X and Y are two random variables, then covariance between them is defined as

$$\begin{aligned}
\text{Cov.}(X, Y) &= E[X - E(X)\} \; \{Y - E(Y)\}] \\
&= E[XY - XE(Y) - YE(X) + E(X)E(Y)] \\
&= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

If X and Y are independent, then $E(XY) = E(X)E(Y)$ and hence in this case

$$\text{cov.}(X, Y) = E(X)\,E(Y) - E(X)\,E(Y) = 0$$

4.4 MOMENT GENERATING FUNCTION

Q. In sampling a large number of Parts. Manufactured by a Machine, the Mean number of defectives in a sample of 20 is 2. Out of 1000 such samples. how many would be expected to contain

(i) at least 3 defective Parts ?

(2) None defective ?

Sol:— Sol:— given Mean $m = 2$

We have $m = np$ $\quad n = 20$ (Parts)

$$2 = 20p \quad \therefore p = \frac{2}{20} = \frac{1}{10} = 0.1$$

$$\therefore q = 1-p = 1-0.1 = 0.9$$

(I) The Probability of at least 3 defective parts in a sample of 20

$$P(r \geq 3) = 1 - P(0) - P(1) - P(2)$$

$$= \left[1 - 20_{C_0}(0.9)^{20} - 20_{C_1}(0.1)^1(0.9)^{19} - 20_{C_2}(0.1)^2(0.9)^{18}\right]$$

$$= 0.323$$

Hence the Required Number of Samples which have at least three defective Parts out of 1000 Such Samples

$$= N \times P(r \geq 3)$$

$$= 1000 \times 0.323 \doteqdot 323$$

(ii) The Probability of None defective Parts in a Sample of 20

$$P(r=0) = 20_{C_0}(0.1)^0(0.9)^{20} = 0.122$$

Hence Required Samples = $1000 \times 0.122 = 122$ Samples

A

Q. A coin is tossed 4 times. What is the probability of getting (i) two heads (ii) at least two heads.

Sol:- The probability of getting a head in a single throw of one coin $p = \frac{1}{2}$

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

No. coin $\boxed{n = 4}$

(I) The probability of getting 2 heads of throwing 4 coins i.e $P(r=2)$ = We have by Binomial distribution

$$P(X=r) = {}^nC_r \, p^r q^{n-r} \quad \text{——— (1)}$$

$$P(r=2) = {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{4-2} = \frac{\lfloor 4}{\lfloor 2 \lfloor 2} \times \frac{1}{4} \times \frac{1}{4}$$

$$= \frac{\cancel{4} \times 3 \times \cancel{2} \times 1}{\cancel{4}} \times \frac{1}{\cancel{4}_{2}} \times \frac{1}{4} = \frac{3}{8} = \boxed{0.375}$$

(II) The probability of getting at least two heads

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

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

$$= \boxed{0.6875}$$

Q. 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.

Sol:- given the Probability of defective Pens is 10%

$$\text{ie } P = 10\% = \frac{1\cancel{0}}{10\cancel{0}} = 0.1$$

$$q = 1 - P = 1 - 0.1 = 0.9 \qquad n = 12 \text{ (Pens in a box)}$$

I) The Probability of exactly two defective Pens out of 12 Pens in a box. We have by Binomial distribution

$$P(X=2) = {}^{n}C_{r}\, p^{r} q^{n-r}$$

$$= {}^{12}C_{2} (0.1)^{2} (0.9)^{10} = 0.2301$$

II) The Probability of at least two defective Pens.

$$P(X \geq 2) = 1 - [P(0) + (P(1)]$$

$$= 1 - [{}^{12}C_{0} (0.1)^{0} (0.9)^{12} + {}^{12}C_{1} (0.1)^{1} (0.9)^{11}]$$

$$= 0.341$$

Q. Fit a binomial distribution to the following data :-

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

sol:- $\boxed{n=5}$ Mean $\bar{x} = \frac{\sum fx}{\sum f} = \frac{284}{100} = 2.84$

mean $m = \bar{x} = np$

$\therefore\ 2.84 = np \quad \therefore\ p = \frac{2.84}{n} = \frac{2.84}{5} = 0.568$

$\therefore\ q = 1-p = 1-0.568 = 0.432$

Hence the binomial distribution to be fitted to the data is $= N(p+q)^n$

$= 100(0.568+0.432)^5$ A

Q. Six dice are thrown 729 times. How many times do you expect at least three dice to show a five or a six?

Sol:- p = the chance of getting 5 or 6 with one a dice $= \frac{2}{6} = \frac{1}{3}$

$q = 1-p = 1-\frac{1}{3} = \frac{2}{3}$ $\boxed{n=6}$ $N = 729$

The Probability to show a 5 or 6 in at least 3 dice.

$= \sum_{x=3}^{6} (P(x)) = P(3) + P(4) + P(5) + P(6)$

where $P(x)$ is the Probability to show 5 or 6,

$= 6_{c_3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^3 + 6_{c_4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^2 + 6_{c_5}\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^1 + 6_{c_6}\left(\frac{1}{3}\right)^6 = 0.3196$

The Required No. $= N \cdot P = 729 \times 0.3196 = \boxed{233}$

Q. out of 800 families with 4 children each how many families would we expected to have.

(i) 2 boys and 2 girls (ii) at least one boy.

(iii) no girl. (iv) at most 2 girls ?

Assume equal probabilities for boys and girls.

Sol:— since the probabilities for boys and girls are equal.

$P =$ probability of having a boy $= \frac{1}{2}$

$\therefore \; q = 1 - P = 1 - \frac{1}{2} = \frac{1}{2}$ $\boxed{n=4}$ $N = 800$ familly

(i) The probability of a family having 2 boys and 2 girls. we have by Binomial distribution

$P(x=2) = n_{C_r} \, p^r q^{n-r}$ —— (1)

$= 4_{C_2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{4-2} = 4_{C_2}\left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{\lfloor 4}{\lfloor 2 \times \lfloor 2} \times \frac{1}{16}$

$\frac{\cancel{4}\times 3\times 2}{\cancel{4}} \times \frac{1}{16} = \frac{3}{8}$

No. of families having 2 boys and 2 girls out of 800 families,

$P(x) = N(P(x)) = \overset{100}{\cancel{800}} \times \frac{3}{\cancel{8}} = \underline{300 \text{ families}}.$

A

(ii) The probability of a family having at least one boys.

$P(x \geq 1) = [1 - P(0)] = 1 - 4_{C_0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4$

$= 1 - \frac{1}{16} = \frac{15}{16} = 0.9375$

The expected No. of families having at least one boys. $= N \times P(r \geq 1)$

$= 800 \times 0.9375 = 750$ families.

(iii) The Probability of a family having no girl. (ie having 4 boys)

ie $P(r=0) = {}^4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{4-0} = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$

The expected No. of families having no. girls ie having 4 boys.

$= 800 \times \frac{1}{16} = 50$ families.

(iv) The Probability of a family having at Most 2 girls?

$$P(r \leq 2) = P(0) + P(1) + P(2)$$

$$= {}^4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 + {}^4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 + {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2$$

$$= \frac{1}{16} + \cancel{4} \times \frac{1}{\cancel{2}} \times \frac{1}{\underset{4}{\cancel{8}}} + \frac{\underline{|4}}{\underline{|2} \times \underline{|2}} \times \frac{1}{4} \times \frac{1}{4}$$

$$\frac{1}{16} + \frac{1}{4} + \frac{\cancel{4} \times 3 \times \cancel{2}}{\underset{2}{\cancel{4} \times \cancel{4} \times 4}} = \frac{1}{16} + \frac{1}{4} + \frac{3}{8}$$

$$\frac{1 + 4 + 3 \times 2}{16} = \frac{11}{16} = 0.6875$$

The expected No. of families having at Most 2 girls. $= 800 \times 0.6875 = 550$ families

Poissian Distribution

It is the limiting form of the Binomial distribution when p (or q) is very small and n is very large so that the average number of success np is finite constant m (say) i.e.

The Probability of r success out of n trials.

$$\boxed{P(r) = \frac{\bar{e}^{m} m^{r}}{\underline{|r}}} \quad , r = 0, 1, 2, 3 - - -$$

Note:- ① m is known as the Parameter of Poissian distribution.

Ex:- ① the Number of defective blades in a Packet of 100 blades.

(ii) the number of misprints on a Page of a book.

(iii) the occurrence of accidents in a factory in a given Period.

Mean, Variance, and standard deviation of Poissian distribution:—

① The Mean or expected value of Poissian distribution is given by

$$\bar{x} = \mu_1' = \sum p_i x_i = \sum P(r) \cdot r$$

$$= \sum_{r=0}^{\infty} \frac{m^{r} \bar{e}^{m}}{\underline{|r}} \cdot r = \sum_{r=0}^{\infty} \frac{m^{r} \bar{e}^{m}}{\underline{|r-1}}$$

$$= m\bar{e}^{m} \sum_{r=1}^{\infty} \frac{m^{r-1}}{\underline{|r-1}} \Rightarrow \bar{e}^{m} \cdot m \cdot e^{m} = m$$

Correlation and Regression

7.1 Introduction

The main objective of many statistical investigations is to make predictions, preferably on the basis of mathematical equations. Problems related to predictions may be treated in two ways (i) by computing the correlation coefficient and (ii) by using regression analysis. The correlation coefficient tells us how strongly two variables are related but it does not give us the magnitude of change of one variable due to other variable. For example the correlation coefficient can tell us whether crime rate and unemployment rate are related or not or whether in computer system the throughput and the degree of multiprogramming are related to each other or not. On the other hand regression model helps us to evaluate the magnitude of change in one variable due to change in other variable. It also helps us to predict the value of one variable for a given value of another variable. For example it estimates the increase in the crime rate due to a particular increase in the unemployment rate or it can predict the food expenditure of a household corresponding to a given income. While the correlation coefficient measures the " closeness", the regression equation is used for prediction or estimation.

This chapter deals with the correlation analysis in which one wishes to find whether a mathematical relationship exists and to measure the strength of such relationship. Here, we also study regression analysis in which the exact nature and form of the mathematical equation of the relation is obtained.

7.2 Bivariate Distribution

The distribution involving one variable is known as univariate distribution. On the other hand if a distribution has two variables it is known as bivariate distribution. For example, for a group of individuals one variable may measure the income while other variable may measure the expenditure and the values form the bivariate distribution.

The simplest way to represent the bivariate data is in the form of a diagram, known as scatter diagram; in which the values (x_i, y_i), $i = 1, 2, \ldots, n$ of the variables are plotted in the xy-plane.

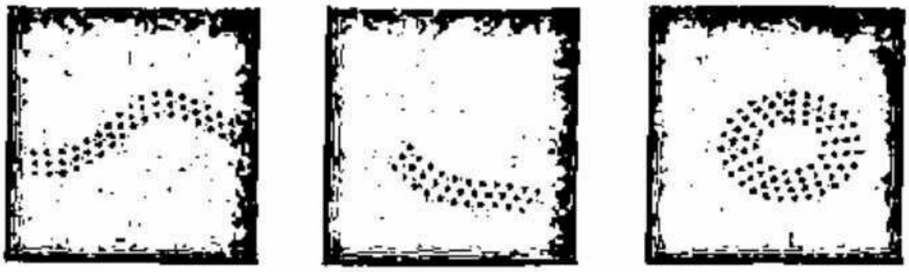


Figure 7.1 Scatter Diagram

7.2.1 Correlation

In a bivariate distribution we may be interested to find out whether there is any correlation between the two variables. The two variables are said to be correlated if change in one variable gives a specific change in the other variable. If the increase (or decrease) in one variable results in the corresponding increase (or decrease) in the other variable, i.e., the two variables deviate in the same direction then the correlation is said to be **direct or positive.** For example, there exists a positive correlation between income and expenditure. If the increase (or decrease) in one variable results in the corresponding decrease (or increase) in the other variable, i.e., the two variables deviate in the opposite directions then the correlation is said to be **diverse or negative.** For example the correlation between volume and pressure of a perfect gas is negative. In a bivariate distribution if the deviation in one variable is followed by the corresponding and proportional deviation (whether positive or negative) in the other then the correlation is said to be **perfect.** For example, $V = L^3$ where V & L are perfectly correlated.

If the correlation between two variables is close to 1, either positive or negative we say variables have a strong positive or negative relationship respectively. If the correlation between two variables is close to zero, whether positive or negative then weak correlation exists.

Some examples of problems which can be referred to as problems of correlation analysis are : the study of relationship between input and output of a waste water treatment plant or a relationship between the tensile strength and the hardness of aluminium or the relationship between impurities in the air and the incidence of a certain disease, where it is assumed that the data points (x_i, y_i) for $i = 1, 2, \ldots, n$ are values of a pair of random variables whose joint density function is $f(x, y)$.

7.2.2 Measure of Correlation: Karl Pearson Coefficient of Correlation

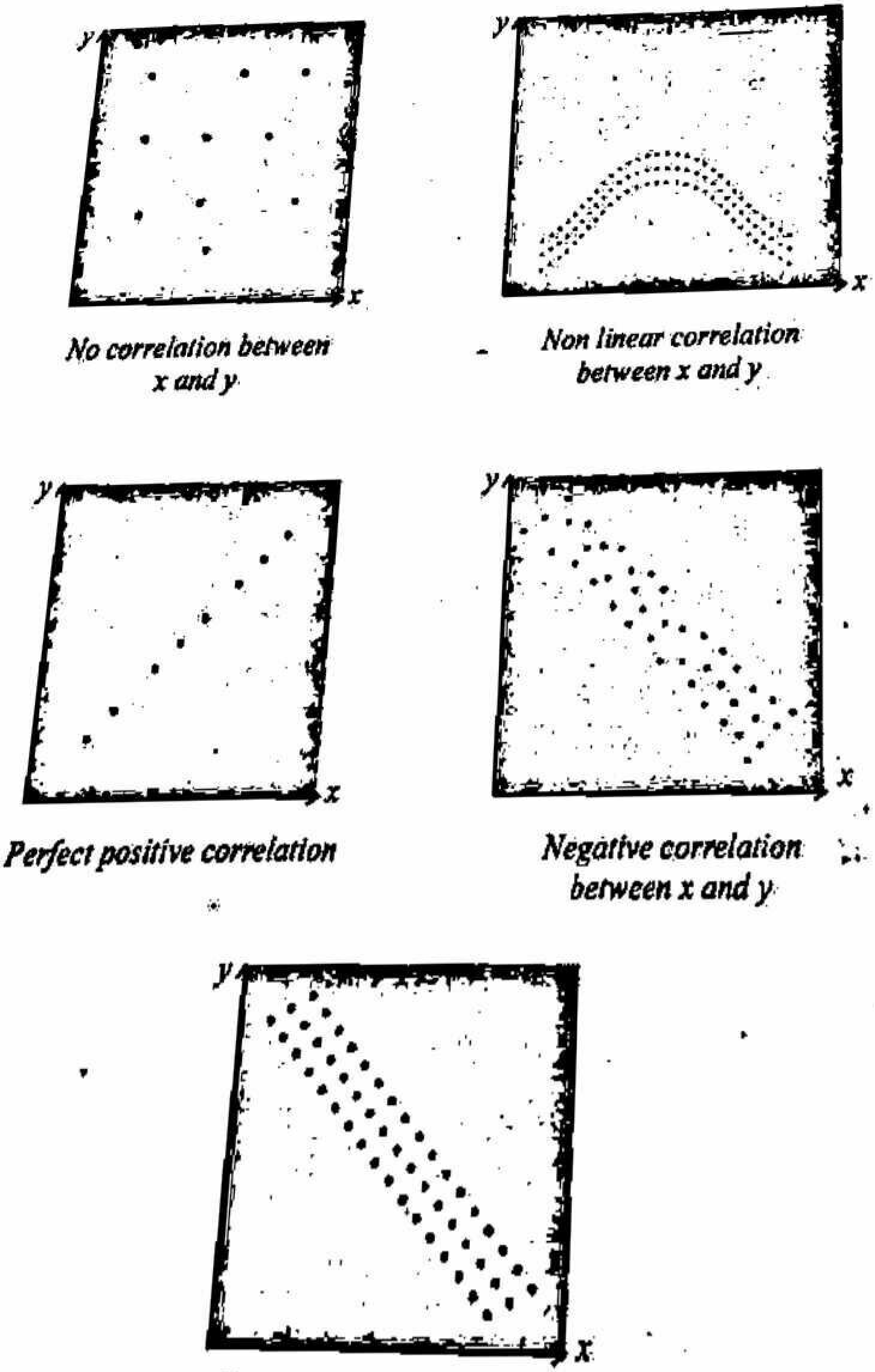
1. Scatter Diagram

As we know the diagram of dots, i.e., the scatter diagram is the simplest way of diagrammatic representation of the given bivariate data. If in this diagram :-

(a) The points are very dense, i.e., very close to each other, then we say that a good amount of correlation exists.

(b) The points are widely scattered then we say that a poor amount of correlation exists.

Some examples are :



No correlation between x and y

Non linear correlation between x and y

Perfect positive correlation

Negative correlation between x and y

Strong negative correlation between x and y

Figure 7.2

But this method is not suitable for a large number of observations. Moreover it does not give the magnitude of correlation.

2. Karl Pearson Coefficient of Correlation

Karl Pearson gave a formula to measure the intensity or degree of linear relationship between two variables, known as correlation coefficient r(X, Y) as

$$r(X,Y) = r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad .....(1)$$

If (x_i, y_i), $i = 1, 2, \ldots, n$ is the bivariate distribution then,

$$Cov(X,Y) = \mu_{11} = E[\{X - E(X)\}\{Y - E(Y)\}]$$

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

Also then

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

and

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

Hence equation (1) ⇒

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

where $\bar{x} = E(X) = \frac{\sum_{i=1}^{n} x_i}{n}$ *and* $\bar{y} = E(Y) = \frac{\sum_{i=1}^{n} y_i}{n}$.

Again we can simplify Cov(X, Y) as :-

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

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

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

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

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

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

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

Similarly

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

Hence substituting these values in equation (2) we get :-

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

This simplified form of r_{XY} is easy for calculations in a bivariate distribution.

We have $Cov(X,Y)=\mu_{11}=\sigma_{XY}$ hence, Karl Pearson's correlation coefficient is also called product-moment correlation coefficient.

Limitations of Correlation Coefficient r_{XY}

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

(ii) Correlation must be used to the data drawn from the same sources. If different 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 also 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.

Remark 1. The scatter diagram for various values of r can be shown as :-

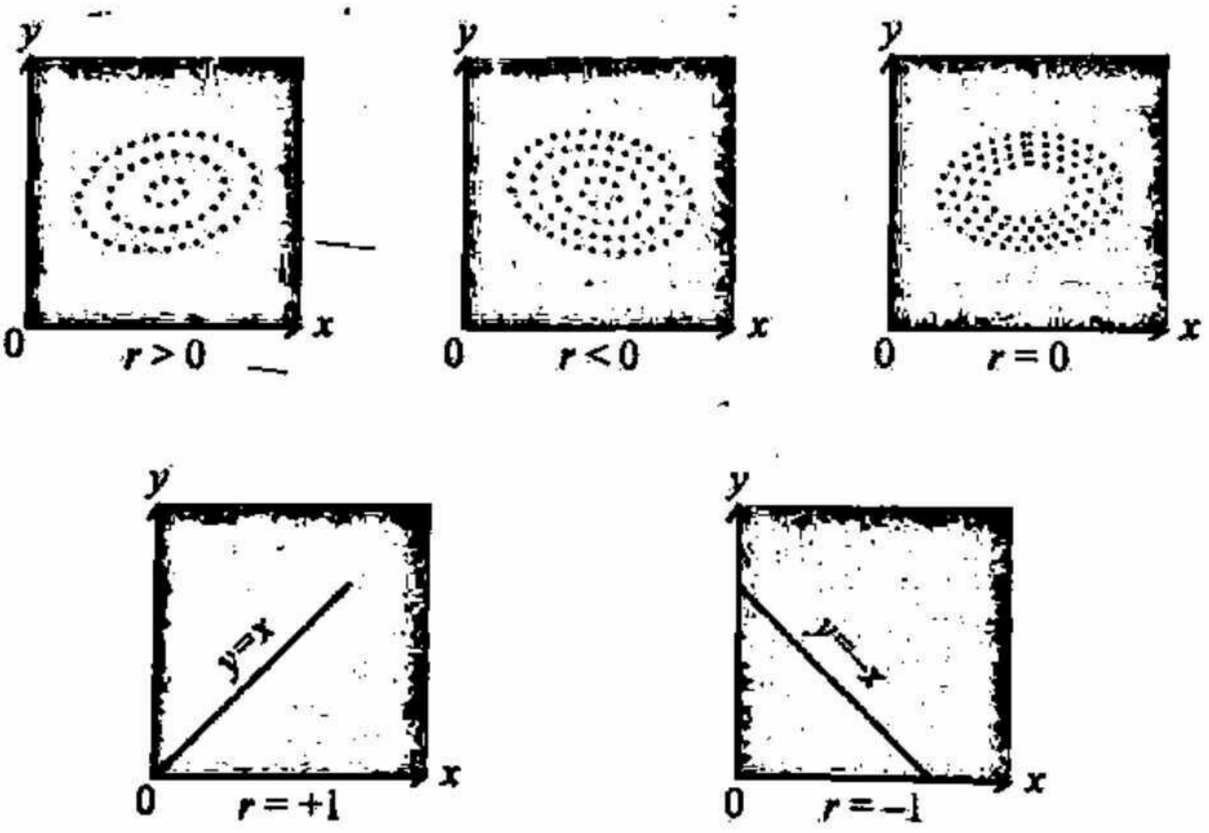


Figure 7.3

Remark 2. The correlation between two variables is said to be simple correlation while between more than two variables is said to be multiple correlation.

Properties of Correlation Coefficient r_{xy}:

(i) Correlation coefficient is independent of change of origin and scale.

Let $U = \dfrac{X-a}{h}$ and $V = \dfrac{Y-b}{k}$

Ex.1 Calculate the correlation coefficient for the following heights (in inches) of fathers (x) and their sons (y):

x :	65	66	67	67	68	69	70	72
y :	67	68	65	68	72	72	69	71

Sol. The correlation coefficient is given by :-

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2}\ \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2}} \quad .....(1)$$

Hence we construct the following table :-

	x	y	x^2	y^2	xy
	65	67	4225	4489	4355
	66	68	4356	4624	4488
	67	65	4489	4225	4355
	67	68	4489	4624	4556
	68	72	4624	5184	4896
	69	72	4761	5184	4968
	70	69	4900	4761	4830
	72	71	5184	5041	5112
Total	544	552	37028	38132	37560

Hence $$\bar{x} = \frac{1}{n}\Sigma x = \frac{544}{8} = 68$$

$$\bar{y} = \frac{1}{n}\Sigma y = \frac{1}{8}\times 552 = 69$$

$\Rightarrow$ $$\sigma_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2} = \sqrt{\left[\frac{37028}{8} - (68)^2\right]} = \sqrt{4.5} = 2.121$$

and $$\sigma_y = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2} = \sqrt{\left[\frac{38132}{8} - (69)^2\right]} = \sqrt{5.5} = 2.345$$

and $$Cov(x,y) = \frac{1}{n}\Sigma xy - \bar{x}\,\bar{y} = \frac{1}{8}\times 37560 - 68\times 69 = 3$$

Therefore equation (1) implies :-

$$r = \frac{3}{2.121\times 2.345} = 0.6032.$$

Aliter

As 'r' is independent of change of origin and scale hence we can simplify the calculati by choosing (may be average values as) arbitrary origin for any one or both of x o As in this example let 68 and 69 be arbitrary origin for x and y respectively. Hence construct the following table

x	y	u=x−68	v=y−69	u^2	v^2	uv
65	67	−3	−2	9	4	6
66	68	−2	−1	4	1	2
67	65	−1	−4	1	16	4
67	68	−1	−1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total		0	0	36	44	24

Now $\bar{u}=\frac{1}{n}\Sigma u=0,\quad \bar{v}=\frac{1}{n}\Sigma v=0,\quad Cov(u,v)=\frac{1}{n}\Sigma uv-\bar{u}\,\bar{v}=\frac{1}{8}\times 24=3$

$$\sigma_u=\sqrt{\frac{1}{n}\Sigma u^2-\bar{u}^2}=\sqrt{\frac{1}{8}\times 36}=\sqrt{4.5}=2.121$$

$$\sigma_v=\sqrt{\frac{1}{n}\Sigma v^2-\bar{v}^2}=\sqrt{\frac{1}{8}\times 44}=\sqrt{5.5}=2.345$$

Hence $$r=\frac{Cov(x,y)}{\sigma_x\,\sigma_y}=\frac{Cov(u,v)}{\sigma_u\,\sigma_v}=\frac{3}{2.121\times 2.345}=0.6032.$$

elation coefficient between the variables x and y, the follo x.3 Calculate the Karl Pearson's coefficient of correlation of the following data :-

x :	25	27	30	35	33	28	36
y :	19	22	27	28	30	23	28

$\bar{x} = \frac{214}{} = 30.5$ and $\bar{y} = \frac{177}{} = 25.2$

7.2.3 Rank Correlation

Although data measured in several cases is numeric (quantitative) in nature but there may be some cases when data turns out to be qualitative or non numeric in nature.

For example, appearance ; beautiful, ugly or efficiency : excellent, good, average, bad or temperature : mild, hot, etc. are some of the cases in which data is qualitative in nature. In such cases, the data is ranked according to the particular characteristic instead of taking numeric measurements on them. Hence here instead of the Pearsonian correlation coefficient its non parametric counterpart developed by Charles Edward Spearman is calculated.

Let us suppose that for a group of n individuals grades or ranks (x_i, y_i) $i = 1, 2, \ldots n$ are given with respect to two characteristics A and B respectively. Then Spearman's rank correlation coefficient for non repeated rank is

$$\rho = 1 - \frac{6\Sigma}{n(n^2-1)} \text{ where } d_i = x_i - y_i .$$

Remark 6. We always have $\Sigma d_i = 0$ as $\Sigma d_i = \Sigma(x_i - y_i) = \Sigma x_i - \Sigma y_i = n(\bar{x} - \bar{y}) = 0$ (as $\bar{x} = \bar{y}$). This can serve as a check for the calculations.

Remark 7 If there are ties, either with respect to characteristic *A* or *B*, substitute for each of the tied observations, the mean of the ranks they jointly occupy.

Remark 8. As Spearman's rank correlation coefficient ρ is same as Pearsonian correlation coefficient between the ranks, hence it can be interpreted in the same way as Karl Pearson's correlation coefficient. Hence $-1 \leq \rho \leq 1$.

Rank Correlation Factor for Repeated Ranks

For repeated ranks, a correction factor is required in the formula. If m is the number of times a rank is repeated then the factor $\frac{m(m^2-1)}{12}$ is to be added to Σd^2.

This correction factor is added for each repeated rank.

Ex.5 The ranks of same 10 students in two subjects A and B are given below :-

Ranks in A	5	2	9	8	1	10	3	4	6	7
Ranks in B	10	5	1	3	8	6	2	7	9	4

Calculate the correlation coefficient

Sol. Here n = 10.

Ranks in $A(x_i)$	5	2	9	8	1	10	3	4	4	7	Total
Ranks in $B(y_i)$	10	5	1	3	8	6	2	7	9	4	
$d_i = x_i - y_i$	–5	–3	8	5	–7	4	1	–3	–3	3	0
d_i^2	25	9	64	25	49	16	1	9	9	9	216

Hence rank correlation coefficient $= \rho = 1 - \dfrac{6\Sigma d^2}{n(n^2-1)} = 1 - \dfrac{6\times 216}{10\times 99} = -0.3091.$

Ex.6 Obtain the rank correlation coefficient for the following data

X :	85	74	85	50	65	78	74	60	74	90
Y :	78	91	78	58	60	72	80	55	68	70

Sol. As we require rank correlation coefficient, hence we have to rank the observations X and Y. First we start with X; 90 being highest value for X receives rank 1. But for next two positions (second and third) we have a tie for value X = 85.. Hence the average rank $\frac{2+3}{2} = \frac{5}{2} = 2.5$ is given to both observations for which value of X is 85, the next rank i.e. the fourth rank goes for the value for which X = 78 and so on.

Similarly we proceed for Y:

Hence we have the following table :-

X	85	74	85	50	65	78	74	60	74	90	Total
Ranks $X(x_i)$	2.5	6	2.5	10	8	4	6	9	6	1	
Y	78	91	78	58	60	~~72~~	80	55	68	70	
Ranks $Y(y_i)$	3.5	1	3.5	9	8	5	2	10	7	6	
$d_i = x_i - y_i$	–1	5	–1	1	0	–1	4	–1	–1	–5	0
d_i^2	1	25	1	1	0	1	16	1	1	25	72

Also for Observation X :-

Rank 2.5 is repeated twice hence its correction factor $= \frac{2(2^2-1)}{2} = \frac{1}{2}$

Rank 6 is repeated twice hence its correction factor $= \frac{3(3^2-1)}{12} = 2$

For Observation Y :

Rank 3.5 is repeated twice, hence its correction factor $= \frac{2(2^2-1)}{12} = \frac{1}{2}$

Hence total correction factor (C.F.) $= \frac{1}{2} + 2 + \frac{1}{2} = 3$

Hence Spearman's rank correlation coefficient is given by

$$\rho = 1 - \frac{6[\Sigma d_i^2 + C.F]}{n(n^2-1)}$$

$$= 1 - \frac{6[72+3]}{10(10^2-1)} = 1 - \frac{6 \times 75}{990}$$

$$= 1 - 0.4545 = 0.5455$$

or $\rho = 0.545$

Sample Problem 2 The ranks of same16 students in mathematics and statistics

Q If x is uniformly distributed with mean 1 and variance $\frac{4}{3}$ find $P(x<0)$

Sol:- for uniform distribution

mean $\mu' = \frac{a+b}{2}$ variance $\sigma^2 = \frac{(a-b)^2}{12}$

$$\frac{a+b}{2} = 1 \quad \therefore a+b = 2 \quad \text{—①}$$

$$\frac{(a-b)^2}{\cancel{12}_{4}} = \frac{4}{\cancel{3}}$$

$$(a-b)^2 = 16$$

$$a-b = \pm 4$$

$$a-b = 4 \quad \text{—②} \qquad a-b = -4 \quad \text{—③}$$

$$\begin{aligned} a+\cancel{b} &= 2 \\ a-\cancel{b} &= 4 \\ \hline 2a &= 6 \end{aligned}$$

$a=3,\ b=-1$

$$\begin{aligned} a+\cancel{b} &= 2 \\ a-\cancel{b} &= -4 \\ \hline 2a &= -2 \end{aligned}$$

$a=-1 \quad b=3$

$$\therefore f(x) = \begin{cases} \frac{1}{b-a} & a<x<b \\ 0 & \text{otherwise.} \end{cases}$$

$$f(x) = \begin{cases} \frac{1}{3-(-1)} & -1<x<3 \\ 0 & o/w \end{cases} = \begin{cases} \frac{1}{4} & -1<x<3 \\ 0 & o/w \end{cases}$$

$$P(x<0) = \int_{-\infty}^{0} f(x)\,dx = \int_{-1}^{0} \frac{1}{4}\,dx = \frac{1}{4}\left[(x)\right]_{-1}^{0} = \frac{1}{4}$$

Ans.

Q ① A Random variable x have an exponential distribution with pdf is given by

$$f(x) = \begin{cases} 2e^{-2x} & x>0 \\ 0 & \text{otherwise } x \le 0 \end{cases}$$

compute the Probability that X is not less Than 3 Also find Mean and S.D, and Prove that coefficient of variation is unity.

Sol:- given $f(x) = \begin{cases} 2e^{-2x} & x>0 \\ 0 & x \le 0 \end{cases}$

$$P(x \not< 3) = P(x \geq 3) = \int_3^{\infty} f(x)\,dx$$

$$= \int_3^{\infty} 2e^{-2x}\,dx = \cancel{2}\left[\frac{e^{-2x}}{-\cancel{2}}\right]_3^{\infty} = e^{-6}$$

$$P(x \not< 3) = e^{-6}.$$

$$\text{Mean} = \mu_1' = \int_0^{\infty} x f(x)\,dx = \int_0^{\infty} x\, 2e^{-2x}\,dx$$

$$2\left[x\frac{e^{-2x}}{-2} - 1\cdot\frac{e^{-2x}}{(-2)^2}\right]_0^{\infty} = 0 + \frac{1}{2} = \frac{1}{2}$$

$$\text{variance} = \sigma^2 = E(x^2) - [E(x)]^2$$

$$= \int_0^{\infty} x^2 f(x)\,dx - (\bar{x})^2$$

$$= \int_0^{\infty} x^2\, 2e^{-2x}\,dx - \frac{1}{4}$$

$$2\left[x^2\frac{e^{-2x}}{-2} - 2x\cdot\frac{e^{-2x}}{(-2)^2} + 2\frac{e^{-2x}}{(-2)^3}\right]_0^{\infty} - \frac{1}{4}$$

$$2\left[0+\frac{2}{8}\right]-\frac{1}{4} \qquad \therefore \frac{1}{2}-\frac{1}{4}=\frac{2-1}{4}=\frac{1}{4}$$

$$S.D=\sqrt{Var.}=\sqrt{\frac{1}{4}}=\frac{1}{2}$$

The coefficient of variation

$$=\frac{Mean}{S.D}=\frac{\frac{1}{2}}{\frac{1}{2}}=1 \quad \text{Proved.}$$

Q. The Income Tax of a man has an exponential distribution with pdf is given by
$$f(x)=\begin{cases}\frac{1}{4}e^{-x/4} & x\geq 0\\ 0 & x<0\end{cases}$$

If income tax is levied at the Rate 5% what is the Probability that his Income exceed ₹ 10000.

sol:- The income Tax on 10000

$$x=\cancel{10000}\times\frac{5}{\cancel{100}}=500\ Rs$$

The Probability that the income exceed ₹ 10000

$$P(X>500)=\int_{500}^{\infty}f(x)\,dx=\int_{500}^{\infty}\frac{1}{4}e^{-x/4}\cdot\cancel{x}\,dx$$

$$=\cancel{\frac{1}{4}}\left[\frac{e^{-x/4}}{-\cancel{\frac{1}{4}}}\right]_{500}^{\infty}=e^{-125}$$

Ans

suppose the life of Mobile battery is exponentially distributed with Parameter $\lambda = 0.001$ days. what is the Probability that a battery will lost more than 1200 days.

ol:- $\lambda = 0.001$, pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$= \begin{cases} 0.001\, e^{-0.001x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Now $P(x \geq 1200) = \int_{1200}^{\infty} f(x)\,dx = \int_{1200}^{\infty} (\lambda e^{-\lambda x})\,dx$

$$= \cancel{\lambda}\left[\frac{e^{-\lambda x}}{-\cancel{\lambda}}\right]_{1200}^{\infty} = 0 - e^{-\lambda \times 1200}$$

$$= e^{-(0.001)\times 1200} = e^{-12}$$

$= 0.301$

Uniform distribution or Rectangular

A continous Random variable X is said to follow a continous uniform or rectangular distribution over an interval (a, b) if its pdf define by

$$f(x) = \begin{cases} K & a<x<b \\ 0 & \text{otherwise.} \end{cases}$$

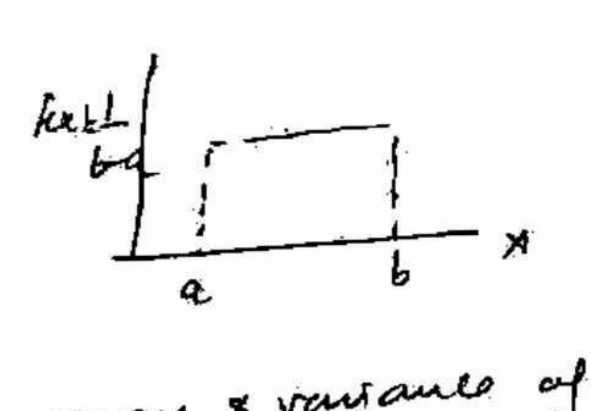
Hence $f(x)$ is pdf

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

$$\int_a^b f(x)\,dx = \int_a^b K\,dx = K\left[x\right]_a^b = K(b-a$$

$$=$$

$$\boxed{K = \frac{1}{b-a}}$$



Mean & variance of uniform distribution

$$\mu_1' = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_a^b x\cdot\frac{1}{b-a}\,dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b$$

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

$$\boxed{\text{mean} = \frac{a+b}{2}}$$

variance :-

$$\sigma^2 = E(x)^2 - [E(x)]^2$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \bar{x}^2$$

$$= \int_a^b x^2 \cancel{k\frac{a+b}{2}}\,dx - \frac{(a+b)^2}{4}$$

$$\int_a^b x^2 \frac{1}{b-a}\,dx - \frac{(a+b)^2}{4}$$

$$\frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b - \frac{(a+b)^2}{4}$$

$$\frac{1}{b-a}\left(\frac{b^3-b^3}{3}\right) - \frac{(a+b)^2}{4}$$

$$= \frac{1}{\cancel{b-a}}\left(\frac{\cancel{(b-a)}(a^2+b^2-ab)}{3}\right) - \frac{(a^2+2ab+b^2)}{4}$$

$$\frac{4a^2+4b^2-4ab-3a^2+6ab+3b^2}{12}$$

$$\sigma^2 = \frac{a^2+b^2-2ab}{12} = \frac{(a-b)^2}{12} =$$

$$\boxed{\text{variance } \sigma^2 = \frac{(a-b)^2}{12}}$$

3.9 NORMAL DISTRIBUTION

It is a theoretical and continuous probability distribution in which the relative frequencies of a continuous variable are distributed according to the normal probability law. In other words, it is a symmetrical distribution in which the frequencies are distributed evenly about the mean of the distribution.

Normal distribution is a limiting form of Binomial distribution under the following conditions.

(i) *n*, the number of trials is infinitely large, *i.e.,* $n \to \infty$.

(ii) *neither p (or q) is very small i.e., p* and *q* are fairly near equally.

A random variable x is said to have a normal distribution with mean 'μ' and standard deviation 'σ' if its probability density function is given by

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

The probability density function with mean zero *i.e.,* $\mu = 0$ and standard deviation σ is given by

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

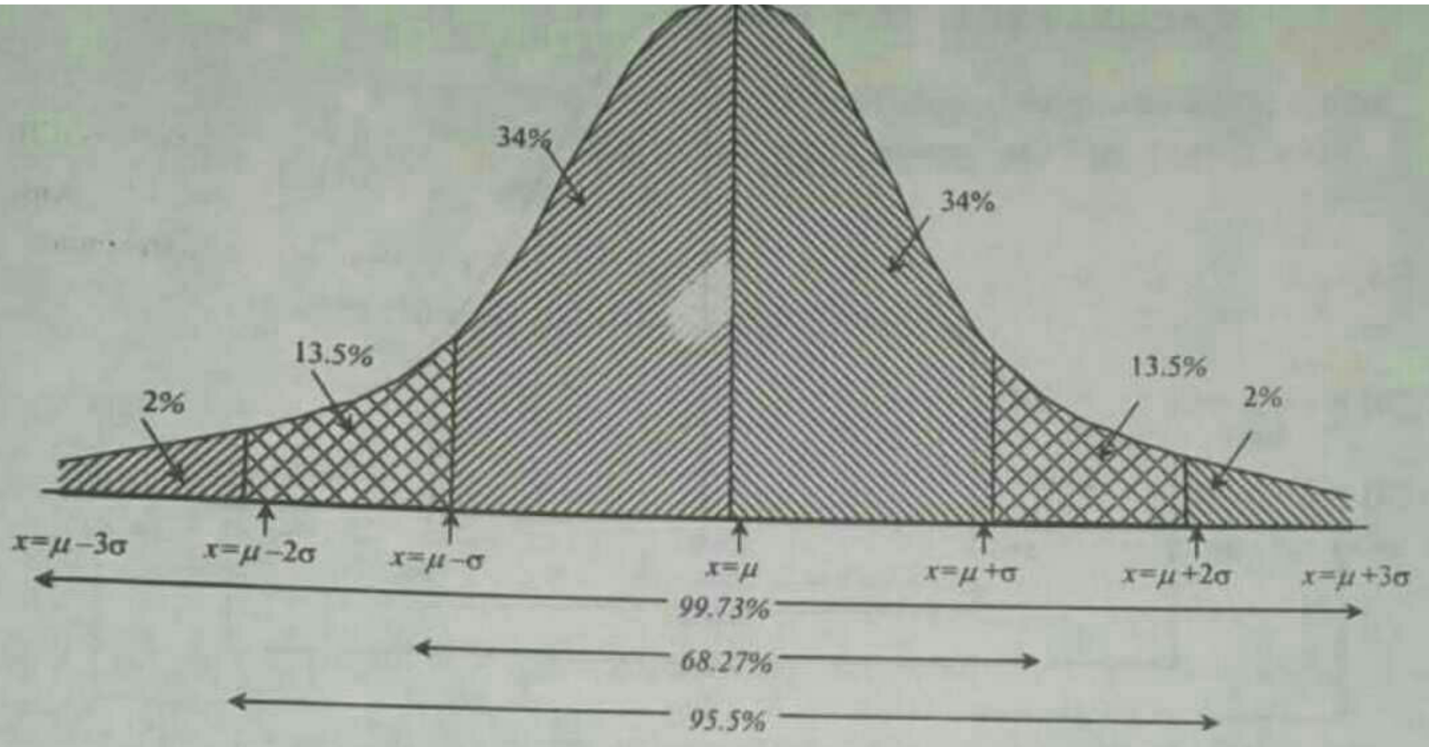
Normal distribution was first discovered by British Mathematician De-Moivre in 1733. Normal distribution is also known as *Gauss an distribution.*

The total area under the normal curve is equal to unity and the *percentage distribution of area under the normal curve* is given below as shown in the figure.

(i) About 68% of the area falls between $\mu - \sigma$ and $\mu + \sigma$.

(ii) About 95.5% of the area falls between $\mu - 2\sigma$ and $\mu + 2\sigma$.

(iii) About 99.7% of the area falls between $\mu - 3\sigma$ and $\mu + 3\sigma$.



5.10 STANDARD NORMAL DISTRIBUTION

A random variable z which has a normal distribution with mean $\mu = 0$ and a standard deviation $\sigma = 1$ is said to have a *standard normal distribution*. Its probability density function is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

It is denoted by $N(0, 1)$. In short, *standard normal variety* is written as *S.N.V.*

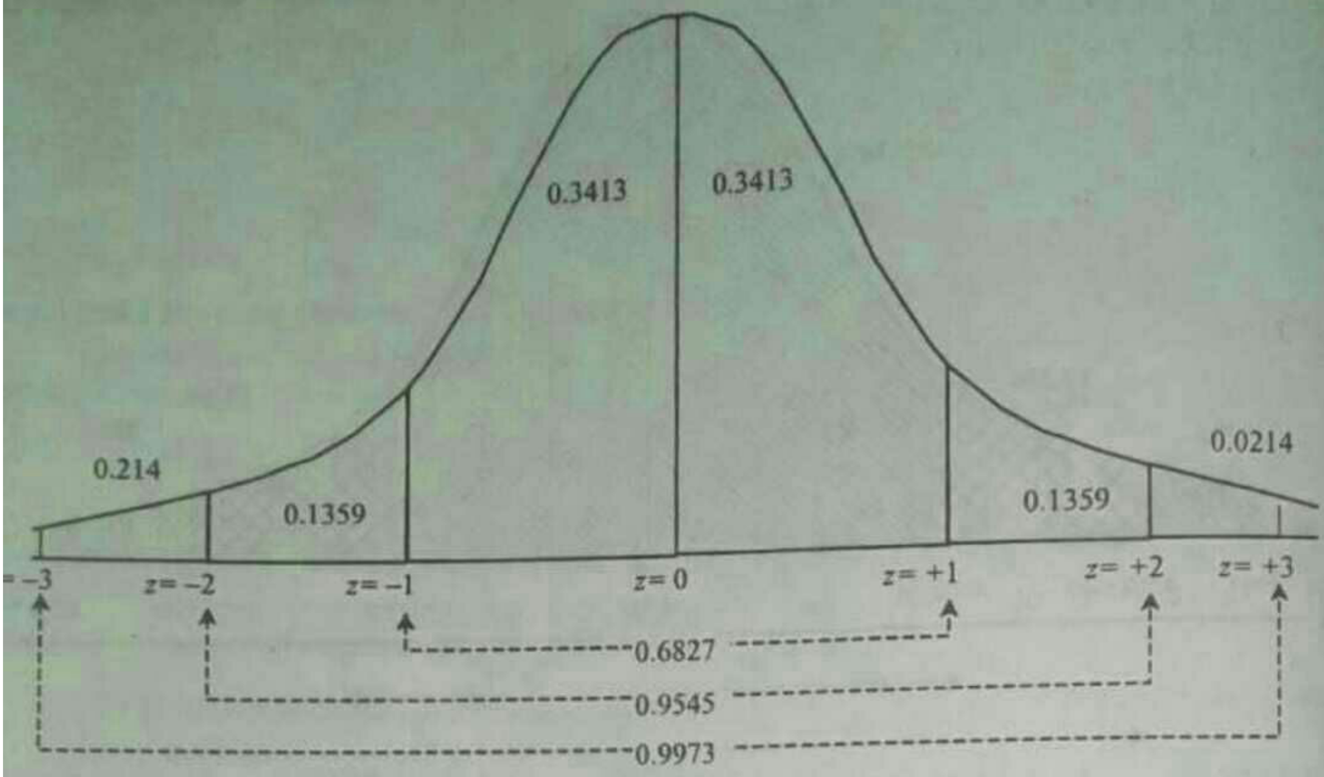
The area under any normal curve is found from the table of a standard normal probability distribution showing the area between the mean and any value of the normally distributed random variable. For a given value of μ and σ, and a specific value, X, of the random variable, the *standardized variety* Z is derived from the following formula :

$$z = \frac{x - \mu}{\sigma}$$

The purpose of standardization of normal distribution is to enable us to make use of the tables of the area of the standard curve $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ for various points along the x-axis.

The standard normal distribution is also known as *Unit Normal Distribution or Z-Distribution.*

The standard normal curve helps us to find the areas within two assigned limits under the curve. The areas between the standard normal curve drawn at two assigned limits a and b will give the proportion of cases for which the values of z lie between a and b. Thus the area between two assigned limits a and b under the standard normal curve will represent the probability that Z will be between a and b. It is denoted by $P(a \le Z \le b)$.



.11 PROPERTIES OF NORMAL CURVE *[RGPV June 2006]*

The normal probability curve with mean μ and standard deviation σ has the following properties :

1. The equation of the curve is

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty \le x \le \infty.$$

and it is bell-shaped. The top of the bell is directly above the mean μ.

2. The curve is symmetrical about the line $x = \mu$ and x ranges from $-\infty$ to $+\infty$.
3. Mean, mode and median coincide at $x = \mu$ as the distribution is symmetrical.
4. X-axis is asymptote to the curve.
5. In Normal distribution: Arithmetic Mean $= \mu$ and Variance $= \sigma^2$
6. The points of inflexion of the curve are at $x = \mu + \sigma$, $x = \mu - \sigma$ and the curve changes from concave to convex at $x = \mu + \sigma$ to $x = \mu - \sigma$.
7. The mean deviation from the mean in normal distribution is equal to 4/5 of its standard deviation.
8. All the odd moments about the mean are zero *i.e.*, $\mu_{2n+1} = 0$.
9. The maximum ordinate lies at the mean *i.e.*, at $x = \mu$.
10. The curve of normal distribution has a single peak, *i.e.*, it is *a unimodal*.
11. The two tails of the curve extend indefinitely and never touch the horizontal line.

5.12 METHOD TO FIND THE PROBABILITY WHEN THE VARIATE IS NORMALLY DISTRIBUTED

Let X be a normal variate with mean μ and standard deviation σ. Suppose we want to find the probability that a randomly selected value for X will lie between a and b *i.e.*, $P(a < X < b)$.

Step 1. Convert X into a standard normal variate by the formula

$$Z = \frac{X-\mu}{\sigma} \qquad \ldots(1)$$

Step 2. Find the limits of Z corresponding to the limits of X.

When $X = a$, then $Z = \dfrac{a-\mu}{\sigma}$ [*Put* $X = a$ *in* (1)]

When $X = b$, then $Z = \dfrac{b-\mu}{\sigma}$ [*Put* $X = b$ *in* (1)]

Thus the limits of Z are $\dfrac{a-\mu}{\sigma}$ to $\dfrac{b-\mu}{\sigma}$, when $X = a$ *to* $X = b$.

Step 3. Thus $P(a < X < b) = P\left(\dfrac{(a-\mu)}{\sigma} < Z < \dfrac{(b-\mu)}{\sigma}\right)$.

Step 4. From the normal table find the probability that Z is between $(a-\mu)/\sigma$ and $(b-\mu)/\sigma$.

Example 5.39 : ***A certain type of wooden beam has a mean breaking strength of 1500 kgs and a standard deviation of 100 kgs. Find the relative frequency of all such beams whose breaking strengths are between 1450 and 1600 kgs.***

Solution. Let X be the breaking strength. Then we are to find $P(1450 < X < 1600)$.

Let X be a normal variety with mean $\mu = 1500$ and standard deviation $\sigma = 100$.

Then standard normal variety $N(0, 1)$:

$$Z = \frac{X-\mu}{\sigma} = \frac{X-1500}{100}.$$

When, $X = 1450$, then $Z = \dfrac{1450-1500}{100} = -0.5$

When $X = 1600$, then $Z = \dfrac{1600-1500}{100} = 1$

Thus $P(1450 < X < 1600) = P(-0.5 < Z < 1)$

$= P(-0.5 < Z < 0) + P(0 < Z < 1)$ [From normal table]

$= 0.1915 + 0.3413 = 0.5328$

$\Rightarrow$ 53% of the beam has the breaking strength between 1450 kgs and 1600 kgs. **Ans.**

Example 5.40 : ***Prove that the points of inflexion of the normal curve are*** $x = \pm\sigma$. ***[RGPV 2001]***

Solution. Let the equation of normal curve be

$$y = y_0 e^{-x^2/2\sigma^2} \qquad \ldots(1)$$

We know that points of inflexion are given by

$$\frac{d^2y}{dx^2} = 0 \text{ and } \frac{d^3y}{dx^3} \neq 0.$$

Example 5.44 : ***For the normal curve***

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}.$$

Find mean and standard deviation.

Solution. Since $f(x)$ is $p.d.f.$, then

$$\text{mean} = \int_{-\infty}^{\infty} x.f(x)\,dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}.x\,dx. \qquad \left[\text{Putting } t = \frac{x-\mu}{\sigma\sqrt{2}} \Rightarrow dx = \sigma\sqrt{2}\,dt.\right]$$

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

$$= \frac{1}{\sigma\sqrt{2\pi}}\left[2\sigma^2 \int_{-\infty}^{\infty} e^{-t^2} t\,dt + \mu\sigma\sqrt{2}\int_{-\infty}^{\infty} e^{-t^2}dt\right]$$

$$\left[\text{Since } \int_{-\infty}^{\infty} e^{-t^2} t\,dt = 0 \text{ and } \int_{-\infty}^{\infty} e^{-t^2}dt = \sqrt{\pi}\right]$$

$$\therefore \quad \text{mean} = \frac{1}{\sigma\sqrt{2\pi}}\left[2\sigma^2 \times 0 + \mu\sigma\sqrt{2}\times\sqrt{\pi}\right]$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \times \mu\sigma\sqrt{2}\times\sqrt{\pi} = \mu$$

Thus mean of normal curve is μ. **Ans.**

Further, by definition of variance

$$\text{Variance} = \int_{-\infty}^{\infty} (x-\text{mean})^2.f(x)\,dx$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2.\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\,dx. \qquad [\because \text{ mean } = \mu]$$

$$\left[\text{Putting, } t = \frac{x-\mu}{\sigma\sqrt{2}} \Rightarrow dx = \sigma\sqrt{2}\,dt.\right]$$

$$\therefore \quad \text{Variance} = \frac{1}{\sigma\sqrt{(2\pi)}} \int_{-\infty}^{\infty} e^{-t^2}\left(t\sigma\sqrt{2}\right)^2 \sigma\sqrt{2}\,dt$$

$$= \frac{2\sigma^3\sqrt{2}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2} t^2 dt = \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left(te^{-t^2}\right)t\,dt.$$

On integrating by parts, we get

$$\text{Variance} = \frac{2\sigma^2}{\sqrt{\pi}}\left[t.\left\{-\frac{1}{2}e^{-t^2}\right\}_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\left(-\frac{1}{2}e^{-t^2}\right)dt\right]$$

$$= \frac{2\sigma^2}{\sqrt{\pi}}\left[0+\frac{1}{2}\int_{-\infty}^{\infty} e^{-t^2}\,dt\right]$$

$$= \frac{2\sigma^2}{\sqrt{\pi}}\times\frac{1}{2}\times\sqrt{\pi} \qquad \left[\because \int_{-\infty}^{\infty} e^{-t^2}dt = \sqrt{\pi}\right]$$

$$= \sigma^2.$$

Hence Variance of normal curve is σ^2. **Ans.**

Example 5.45 : The mean deviation from the mean of the normal distribution is $\frac{4}{5}$ times its standard deviation.

Or

Find the mean deviation from mean for normal distribution. *[RGPV June 2009, Feb. 2010]*

Solution: By Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2} \qquad ...(1)$$

We know that, mean deviation from mean 'μ':

$$\text{M.D.} = \int_{-\infty}^{\infty} |x-\mu|\, f(x)\,dx$$

$$= \int_{-\infty}^{\infty} |x-\mu|.\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2} \qquad [\because \text{by} 1]$$

$$\left[\text{Put, } t = \frac{x-\mu}{\sigma\sqrt{2}} \Rightarrow dt = \frac{dx}{\sigma\sqrt{2}} \Rightarrow dx = \sigma\sqrt{2}dt\right]$$

$$\therefore \quad \text{M.D.} = \int_{-\infty}^{\infty} |\sqrt{2}\,\sigma t|.\frac{1}{\sigma\sqrt{2\pi}}\,e^{-t^2}.\sigma\sqrt{2}\,dt$$

$$= \frac{\sigma}{\sqrt{\pi}}.\sqrt{2}\int_{-\infty}^{\infty} |t|e^{-t^2}\,dt$$

$$= \frac{\sigma\sqrt{2}}{\sqrt{\pi}}.\left(2\int_0^{\infty} t\,e^{-t^2}\,dt\right) \qquad \left[\because \int_{\infty}^{\infty} \text{even } f^n = 2\int_0^{\infty}\right]$$

Put $u = t^2 \Rightarrow\ du = 2t\,dt\ \Rightarrow t dt = \frac{du}{2}$

$$\therefore \quad \text{M.D.} = \frac{2\sigma\sqrt{2}}{\sqrt{\pi}}\int_0^{\infty} e^{-u}\,\frac{du}{2} = \frac{\sigma}{\sqrt{2\pi}}\left[\frac{e^{-u}}{-1}\right]_0^{\infty}$$

$$= \frac{-2\sigma}{\sqrt{2\pi}}[0-1] = \sigma\sqrt{\frac{2}{\pi}}$$

$$= (0.8)\sigma \text{ approx}$$

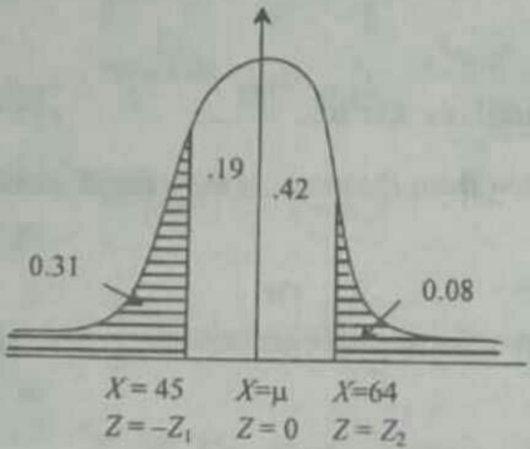
$$= \frac{4\sigma}{5} \text{ approx}$$

Proved.

Example 5.46 : In a normal distribution 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution. ***[RGPV June 2002 & June 2007]***

Solution. 31% of items are under 45⇒ Area to the left is 0.31. But area right from this point is (0.50–0.31) = 0.19. (See figure)

Let X be the random variety, which is normally distributed with mean μ and standard deviation σ.



Then, $Z = \dfrac{X-\mu}{\sigma}$ is a standard normal variable (S.N.V.)

The S.N.V. corresponding to $X = 45$ and $X = 64$ are as below :

When $X = 45$, then $Z = \dfrac{45-\mu}{\sigma} = -Z_1$. (Say) ...(1)

When $X = 64$, then $Z = \dfrac{64-\mu}{\sigma} = Z_2$. (Say) ...(2)

From the figure it is obvious that

$P(0 < Z < Z_2) = 0.42 \Rightarrow Z_2 = 1.405$ [From the normal table].

and $P(-Z_1 < Z < 0) = 0.19 \Rightarrow P(0 < Z < Z_1) = 0.19.$ [by symmetry]

$\Rightarrow \quad Z_1 = 0.496$ [From the normal table].

Substituting the values of Z_1 and Z_2 in (1) and (2), we get

$$\frac{45-\mu}{\sigma} = -0.496 \Rightarrow 45-\mu = -0.496\sigma \qquad ...(3)$$

$$\frac{64-\mu}{\sigma} = 1.405 \Rightarrow 64-\mu = 1.405\sigma \qquad ...(4)$$

Solving (3) and (4), we get

$\sigma = 10$, $\mu = 49.96 \cong 50$ (approx.) *i.e.*, S.D. = 10 and mean = 50.

Ans.

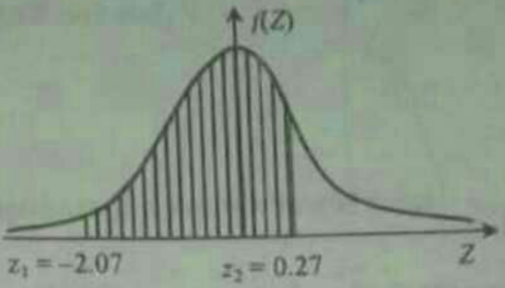
xample 5.47 : The mean height of 500 students is 151 cm, and the standard deviation is 15 cm. Assuming that the heights are normally distributed, find how many students have heights between 120 and 155 cm. ? ***[RGPV Dec. 2003]***

Solution. No. of students $= 500 \quad \therefore N = 500$

Mean, $\mu = 151\,cm$, and $\sigma = 15$

By standard normal variable $z = \dfrac{x-\mu}{\sigma}$.

When $x_1 = 120\,cm$,



Standard normal variable $z_1 = \dfrac{x_1-\mu}{\sigma} = \dfrac{120-151}{15} = \dfrac{-31}{15} = -2.07$.

When $x_2 = 155\,cm$,

Standard normal variable $z_2 = \dfrac{x_2-\mu}{\sigma} = \dfrac{155-151}{15} = \dfrac{4}{15} = 0.27$.

$$\begin{aligned}\therefore \quad P(120 < x < 155) &= P(-2.07 < z < 0.27)\\ &= P(-2.07 \le z \le 0) + P(0 \le z \le 0.27)\\ &= P(0 \le z \le 2.07) + P(0 \le z \le 0.27)\\ &= 0.4808 + 0.1085 \qquad \text{[by normal table]}\\ &= 0.5892.\end{aligned}$$

$\therefore$ The required no. of students $= 0.5892 \times 500 = 294$. **Ans.**

Example 5.48 : The distribution of weekly wages of 500 workers in a factory is approximately normal with the means and standard deviation of Rs.75 and Rs.15. Find the number of workers who receive weekly wages:

(i) More than Rs. 90

(ii) Less than Rs. 45.

Solution: Given: No. of workers $= 500$ *i.e.*, $N = 500$

Mean, $\mu = 75$ and *S.D.* $\sigma = 15$

By standard normal variable $z = \dfrac{x-\mu}{\sigma}$ ...(1)

(i) When, $x = 90$: Then standard normal variable

$$z = \frac{x-\mu}{\sigma} = \frac{90-75}{15} = \frac{15}{15} = 1$$

$$\begin{aligned}\therefore \quad \text{P (more than Rs. 90)} &= P(x > 90)\\ &= P(z > 1)\\ &= 0.5 - P(0 < z < 1)\\ &= 0.5 - 0.3413\\ &= 0.1587\end{aligned}$$

Hence number of workers who receive weekly wages more than Rs. 90

$$= 0.1587 \times 500$$
$$= 79.35 \approx 79 \text{ workers}$$

(ii) When, $x = 45$: Then $z = \dfrac{x-\mu}{\sigma} = \dfrac{45-75}{15} = -2$

$$\therefore \qquad \text{P (less than Rs. 45)} = P(x<45) = P(z<-2)$$
$$= 0.5 - P(-2<z<0)$$
$$= 0.5 - 0.4772$$
$$= 0.0228$$

Hence No. of workers who receive wages less than Rs. 45

$$= 0.0228 \times 500 = 11.4 \approx 11 \text{ workers}$$

Ans.

Example 5.49 : ***A sample of 100 dry battery cells tested to find the length of life produced the following results***

Mean $\bar{x} = 12$ hours, standard deviation $\sigma = 3$ hours.

Assuming the data to be normally distributed, what percentage of battery cells are expected to have life

(i) More than 15 hours ***(ii) Less than 6 hours***

(iii) Between 10 and 14 hours ? *[RGPV Feb. 2005]*

Solution. Here x denotes the length of life of dry battery cells.

The S.N.V. : $z = \dfrac{x-\bar{x}}{\sigma} - \dfrac{x-12}{3}$.

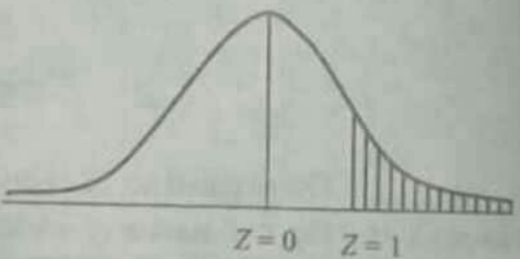
(i) When $x = 15$, then $z = 1$.

$\therefore$ $P(\text{more then } 15) = P(x>15) = P(z>1)$

$= P(0<z<\infty) - P(0<z<1)$

$= 0.5 - 0.3413 = 0.1587 = 15.87\%$.

Ans.

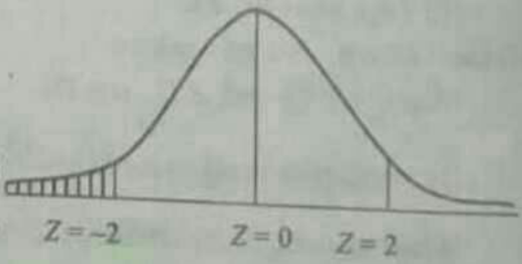


(ii) When $x = 6$, then $z = \dfrac{6-12}{3} = -2$

$\therefore$ $P(\text{less then } 6) = P(x<6) = P(z<-2)$

$= P(z>2) = P(0<z<\infty) - P(0<z<2)$

$= 0.5 - 0.4772 = 0.0228 = 2.28\%$. Ans.



(iii) When $x = 10$, then $z = \dfrac{10-12}{3} = -0.67$.

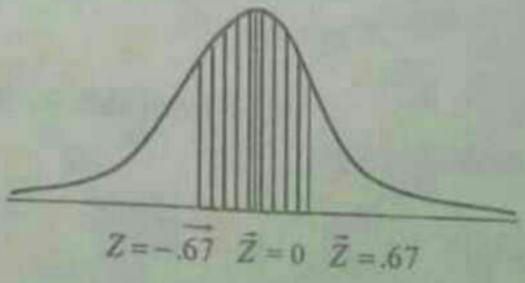
when $x = 14$ then $z = \dfrac{14-12}{3} = 0.67$.

$\therefore$ $P(\text{Between}10\&14)$

$= P(10<x<14) = P(-0.67<z<0.67)$

$= 2P(0<z<0.67) = 2 \times 0.2487$

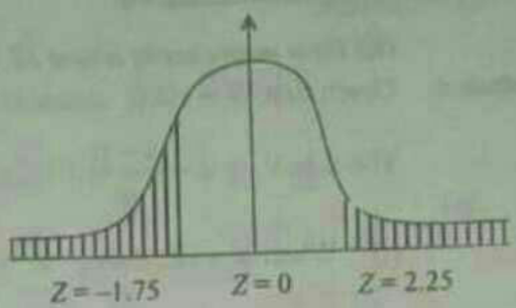
$= 0.4974 = 49.74\%$.



Ans.

Example 5.50 : ***Assuming that the diameters of 1000 brass plugs taken consecutively from a machine from a normal distribution with mean 0.7515 cm. and standard deviation 0.0020 cm., how many of the plugs are likely to be rejected, if the approved diameter is 0.752 ± 0.004 cm.***

[RGPV Dec. 2002]



Solution. Given range of diameter are :

$x_1 = 0.752 - 0.004 = 0.748\,cm.$ and

$x_2 = 0.752 + 0.004 = 0.756\,cm.$

Mean $\mu = 0.7515$ and S.D. $\sigma = 0.0020\,cm.$

$\because$ The S.N.V. is $z = \dfrac{x-\mu}{\sigma}$

$\therefore$ At $x_1 = 0.748$, then $z_1 = \dfrac{0.748 - 0.7515}{0.002} = -1.75.$

Also at $x_2 = 0.756$, then $z_2 = \dfrac{0.756 - 0.7515}{0.002} = 2.25.$

$$\begin{aligned}\therefore \quad P(x_1 < x < x_2) &= P(z_1 < z < z_2)\\ &= P(-1.75 < z < 2.25)\\ &= P(0 < z < -1.75) + P(0 < z < 2.25)\\ &= 0.4599 + 0.4878\\ &= 0.9477.\end{aligned}$$

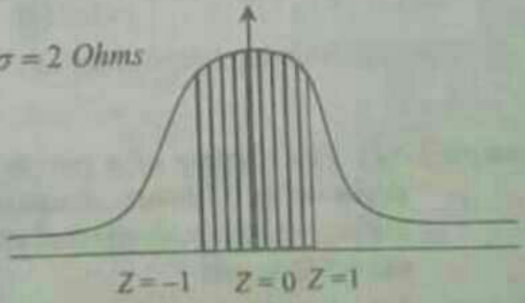
[From normal table]

$\therefore$ Number of plugs likely to be rejected $= 1000(1 - 0.9477)$

$= 1000 \times 0.0523$ [$\because$ Shaded area = (1−0.9477)]

$= 52.3 \approx 52.$ **Ans.**

Example 5.51 : A manufacturer knows from experience that the resistance of resistors be produces is normal with mean $\mu = 100$ Ohms and standard deviation $\sigma = 2$ Ohms. What percentage of resistors will have resistance between 98 Ohms and 102 Ohms ?



Solution. Given that mean $\mu = 100\,Ohms$, standard deviation $\sigma = 2$ *Ohms*

By standard normal variable is $z = \dfrac{x-\mu}{\sigma}.$

When $x = 98$, then $z_1 = \dfrac{98-100}{2} = -1.$

When $x = 102$, then $z_2 = \dfrac{102-100}{2} = 1.$

$$\begin{aligned}\therefore \quad P(98 < x < 102) &= P(z_1 = -1 < z < z_2 = 1)\\ &= P(-1 < z < 1)\\ &= P(-1 < z < 0) + P(0 < z < 1)\\ &= 0.3413 + 0.3413 = 0.6826.\end{aligned}$$

Thus, the percentage of resistors having resistance between 98 *ohms* and 102 *ohms*.

$= 0.6826 \times 100 = 68.26\%$ **Ans.**

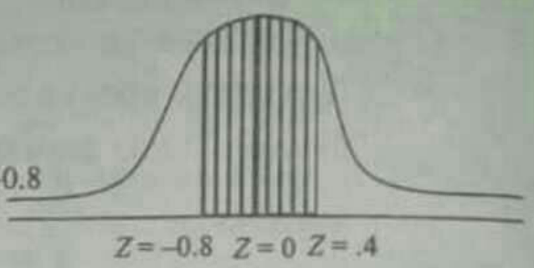
Example 5.52 : In a sample of 1000 cases, the mean of a certain test is 14 and standard deviation is 2.5. Assuming the distribution to be normal find

(i) How many students score between 12 and 15 ?

(ii) How many score above 18 ?

Solution. Given that $N = 1000$, *mean* $\mu = 14$, $\sigma = 2.5$.

The S.N.V. is $z = \dfrac{x - \mu}{\sigma}$



(i) When $x = 12$, then, $z_1 = \dfrac{x-\mu}{\sigma} = \dfrac{12-14}{2.5} = -0.8$

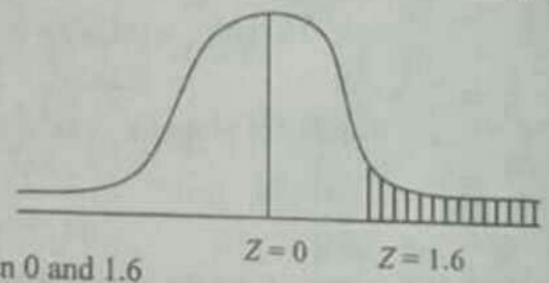
When $x = 15$, then, $z_2 = \dfrac{15-14}{2.5} = 0.4$

$$\therefore \quad P(12 < x < 15) = P(-0.8 < z < 0.4)$$
$$= P(-0.8 < z < 0) + P(0 < z < 0.4)$$
$$= 0.2881 + 0.1554 \qquad \text{[From normal table]}$$
$$= 0.4435.$$

$\therefore$ The required number of students $= 1000 \times 0.4435$

$= 443.5 \approx 444.$ **Ans.**

(ii) When $x = 18$, then, $z = \dfrac{18-14}{2.5} = 1.6$



$\therefore$ $P(\text{score above } 18) = P(x > 18) = P(z > 1.6)$

= Left area – Area between 0 and 1.6

$= 0.5 - P(0 < z < 1.6)$

$= 0.5 - 0.4452$ [From normal table]

$= 0.0548.$

$\therefore$ The required number of students $= 1000 \times 0.0548$

$= 54.8 \approx 55.$ **Ans.**

Example 5.53 : The lifetime of a certain kind of battery has a mean of 300 hours and a standard deviation of 35 hours. Assuming that the distribution of life times, which are measured to the nearest hour, is normal, find the percentage of batteries, which have lifetime of more than 370 hours.

Solution. Let x be a random normal variate measuring the life time of batteries

Here mean $\mu = 300, \sigma = 35$

The S.N.V. is $z = \dfrac{x-\mu}{\sigma} = \dfrac{x-300}{35}$

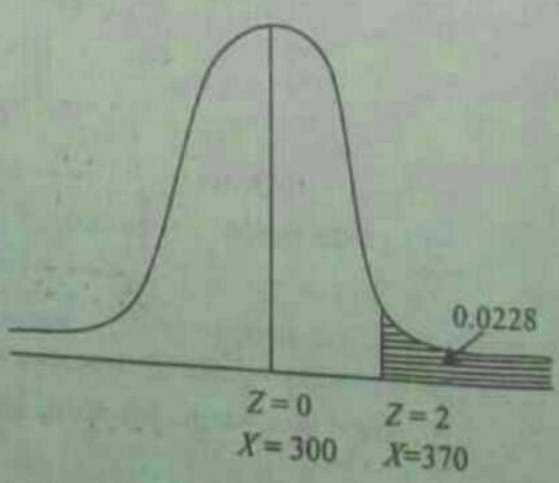
When $x = 370$, then $z = \dfrac{370-300}{35} = 2.$

$\therefore$ P (More than 370)

$= P(x > 370) = P(z > 2)$

= left area – Area between $z = 0$ and $z = 2$

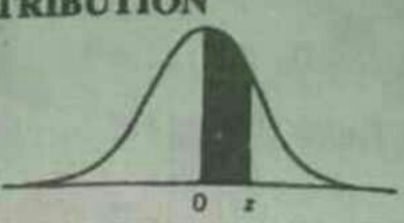
$z = 0$*and* $z = 2$



TABLE

AREA OF A STANDARD NORMAL DISTRIBUTION

An entry in the table is the proportion under the entire curve which is between z = 0 and a postive value of z. Area for negative values of z are obtained by symmetry.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.8832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4415	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	..4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4841	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

The Poisson distribution was discovered by a French Mathematics Simen Denis Poisson in 1837. It is a discrete distribution and is very widely used. Poisson distribution is a limiting form of the Binomial distribution in which n, the number of trials, becomes very large and p, the probability of the success of the event, is very-very small such that mean $m = np$ is a finite quantity.

$\therefore$ Probability of r-successes.

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

Also theoretical or expected frequencies is

$$f(X=r)=N.\frac{e^{-m}m^{r}}{r!},\text{ where } m=\text{mean and } N \text{ is number of trials.}$$

The following are the statistical measures of the Poisson distribution.

(i) Mean $= np$ ro m.

(ii) Variance $= m$ or np.

(iii) Standard Deviation $\sigma = \sqrt{m}$.

(iv) Moment measure of skewness $(\gamma_1)=\frac{1}{\sqrt{m}}$.

(v) Moment measure of kurtosis $(\gamma_2)=\frac{1}{m}$.

Some examples of Poisson distribution :

1. The number of deaths in a city in one year by a rare disease.
2. The number of printing mistakes in each page of the first proof of a book.
3. The number of defective screws per box of 100 screws.

5.7 CONDITIONS UNDER WHICH POISSON DISTRIBUTION IS USED

1. The random variable $X=r$ should be discrete *i.e.*, $r=0,1,2,\ldots\ldots n$. here n is large.
2. The happening of the events must be of two alternatives such as success and failure.
3. It is applicable in those cases when the number of trials n is very large and probability of success p is very small but the mean m is finite.
4. p should be very small (close to zero). If $p\to 0$, then the Poisson distribution is *J-shaped* and *unimodal*.

Example 5.23 : Prove that Poisson distribution as a limiting case of Binomial distribution, when $n\to\infty$.

Solution. Binomial distribution

$$P(X=r)={}^nC_r\,q^{n-r}p^r=\frac{n!}{r!(n-r)!}(1-p)^{n-r}p^r \qquad [\because q=1-p]$$

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

$$=\frac{n(n-1)(n-2)\ldots(n-r+1)}{r!}\times\left(1-\frac{m}{n}\right)^{n-r}\times\left(\frac{m}{n}\right)^r \qquad \left[\because np=m \ \therefore p=\frac{m}{n}\right]$$

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

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

$$=\frac{m}{r!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\ldots\left(1-\frac{r-1}{n}\right)\times\frac{\left[\left(1-\frac{m}{n}\right)^{-\frac{n}{m}}\right]^{-m}}{\left(1-\frac{m}{n}\right)^r} \qquad \ldots(1)$$

Since n is very large so that

$\therefore$ Lim $n \to \infty \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\dots\left(1-\frac{r-1}{n}\right)=(1-0)\ (1-0)\dots\dots(1-0)=1$ and

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

Then (1) becomes :

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

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

which is probability function of Poisson distribution. **Proved.**

5.8 MEAN VARIANCE AND STANDARD DEVIATION OF THE POISSON DISTRIBUTION

For the Poisson distribution

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

We know that

$$\text{mean } =\mu_1' \qquad \dots(1)$$

and $\quad$ variance $=\mu_2'-(\mu_1')^2,$

where μ_1' and μ_2' are first and second moment about origin. ...(2)

For $\quad r=0,1,2,\dots,\infty$ *first moment about origin.*

$$\mu_1'=\sum_{r=0}^{\infty} r\ P(r)$$

$$=\sum_{r=0}^{\infty} r.\frac{e^{-m}\,m^r}{r!} \qquad \dots(3) \qquad [m = \text{mean}]$$

$$=e^{-m}\sum_{r=0}^{\infty}\ \frac{m^r}{(r-1)!}=e^{-m}\left[m+\frac{m^2}{1!}+\frac{m^3}{2!}+\dots\dots\right]$$

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

or $\quad \mu_1'=me^{-m}.e^{m}=m \qquad \dots(4)$

[RGPV Feb. 2006 & June 2008 (N)]

From (1) mean $=\mu_1'=m.$

Now, *second moment about origin,*

$$\mu_2'=\sum_{r=0}^{\infty} r^2\ P(r) \qquad [\text{Since } r^2=r(r-1)+r\,]$$

$$=\sum_{r=0}^{\infty} r(r-1)\frac{e^{-m} m^r}{r!}+\sum_{r=0}^{\infty} r\frac{e^{-m}m^r}{r!} \qquad [m = \text{mean}]$$

$$=e^{-m}\sum_{r=2}^{\infty} \frac{m^r}{(r-2)!}+\mu_1{}'=e^{-m}\left[m^2+\frac{m^3}{1!}+\frac{m^4}{2!}+\ldots\ldots\right]+m \qquad [\text{by (3) and (4)}]$$

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

$$\therefore \qquad \mu_2{}'=m^2+m \qquad \ldots(5)$$

From (2), variance

$$=\mu_2{}'-(\mu_1{}')^2$$

$$=m^2+m-(m)^2 \qquad [\text{by (4) \& (5)}]$$

$$=m.$$

Thus, variance $= m$. *[RGPV Feb. 2006 & June 2008 (N)]*

Further, standard deviation, $\sigma=\sqrt{\text{variance}}=\sqrt{m}.$

Example 5.24 : Show that in a Poisson distribution with unit mean, the mean deviation about the mean, is 2/e times the standard deviation.

Solution. By Poisson distribution

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

Given that mean $m=1$.

We know that mean deviation about mean, $M.D.=\dfrac{\sum_{r=0}^{\infty} f|x-m|}{N}$

$$=\frac{\sum_{r=0}^{\infty} f|r-1|}{\Sigma f} \qquad [\because m=1, N=\Sigma f \ and \ x=r] \qquad \ldots(1)$$

Here $\qquad f=P(r)=e^{-m}\dfrac{m^r}{r!}=e^{-1}\dfrac{(1)^r}{r!}=\dfrac{e^{-1}}{r!} \qquad [\because m=1]$

and $\qquad \Sigma f=\sum_{r=0}^{\infty} P(r)=1 \qquad [\because \text{ total frequencies} = 1]$

Hence (1) becomes :

$$\therefore \qquad M.D.=\sum_{r=0}^{\infty} \frac{e^{-1}|r-1|}{r!}$$

$$=e^{-1}\left[\frac{|0-1|}{0!}+\frac{|1-1|}{1!}+\frac{|2-1|}{2!}+\frac{|3-1|}{3!}+\ldots\ldots\ldots\right]$$

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

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

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

$$=e^{-1}[1+1]=2e^{-1}=\frac{2}{e}. \qquad \ldots(2)$$

Since standard deviation of Poisson deviation $S.D.=\sqrt{m}=\sqrt{1}=1.$ $[\because m=1]$

Hence (2) becomes: $M.D.=\frac{2}{e}.1=\frac{2}{e}(S.D.).$ **Proved.**

Example 5.25 : For the Poisson distribution, prove that

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

Which is known as *recurrence relation* for Poisson distribution.

Solution. By Poisson distribution

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

replacing $r\rightarrow r+1$.

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

Now, $\frac{P(r+1)}{P(r)}=\frac{e^{-m}m^{r+1}}{(r+1)!}\times\frac{r!}{e^{-m}m^r}=m.\frac{r!}{(r+1)r!}=\frac{m}{r+1}.$

Hence $P(r+1)=\frac{m}{(r+1)}P(r).$ **Proved.**

Example 5.26 : Find the probability that at must 5 defective fuses will be found in a box of 200 fuses, if experience shows that 2 percent of such fuses are defective.

olution. Given : $p=2\%=\frac{2}{100}=.02,\ n=200$

$\therefore$ *mean* $m=np\Rightarrow m=200\times.02\Rightarrow m=4.$

Since n is large so that using Poisson distribution

$$P(r)=\frac{e^{-m}m^r}{r!},\ r=0,1,2,\ldots\ldots,200. \qquad \ldots(1$$

$\therefore P$ (at most 5 defective) $=P(r\leq 5)$

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

$$=e^{-4}\left[1+\frac{4}{1!}+\frac{4^2}{2!}+\frac{4^3}{3!}+\frac{4^4}{4!}+\frac{4^5}{5!}\right] \quad [\because \text{ by } (1)]$$

$$=0.0183[1+4+8+10.6667+10.6667+8.5333]$$

$$=0.0183\times 42.8667=0.7845.$$ **Ans.**

Example 5.27 : *In a certain factory turning razor blades, there is a small chance of 0.002 for any blade to be defective. The blades are in packets of 10. Use Poisson's distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades respectively in a consignment of 50,000 packets.* ***[RGPV June 2006]***

Solution. Given, $p=0.002$, $n=10$, $N=50{,}000$

$\therefore \quad m=np = 10\times 0.002=0.02$.

By Poisson distribution

$$P(X=r) = \frac{e^{-m}\,m^r}{r!}, \quad r=0,1,3,\ldots\ldots 10$$

(i) $P(\text{no defictive})=P(r=0) = e^{-0.02}\,\dfrac{(.02)^0}{0!}=0.9802$.

Hence number of packets containing no defective blades

$$=N.P(r=0)=50{,}000\times 0.9802=49010.$$ **Ans.**

(ii) P (one defective) $=P(r=1)=e^{-0.02}.\dfrac{(0.02)^1}{1!}=0.019604$

$\therefore$ Number of packets containing one defective blade

$$=N.P(r=1)=50{,}000\times 0.019604=980.$$ **Ans.**

(iii) P (two defective) $=P(r=2)=e^{-0.02}\,\dfrac{(0.02)^2}{2!}=0.00019604$.

$\therefore$ Number of packets containing two defective blades

$$=N.P(r=2)=50{,}000\times 0.00019604=9.8\approx 10.$$ **Ans.**

Example 5.28 : *If the probability that an individual suffers a bad reaction from a certain injection is 0.001, determine the probability that out of 2000 individuals*

(i) exactly 3 *(ii) more than 2 individuals*

(iii) none *(iv) more that 1 individuals*

will suffer a bad reaction. ***[RGPV June 2003]***

Solution. Given $p=0.001$ (which is very small)

$n=2000$ (which is large), then $m=np \Rightarrow m=2000\times 0.001=2$

Using Poisson distribution :

$$P(X=r)=e^{-m}\frac{m^r}{r!}, \qquad r=0,1,2,\ldots\ldots\ldots 2000.$$

(i) P (exactly 3 suffer a bad reaction) = P (r = 3)

$$=\frac{m^3e^{-m}}{3!}=\frac{8e^{-2}}{6}=\frac{4}{3e^2}=0.180.$$ **Ans.**

(ii) P (more than 2 suffer a bad reaction) $=P(r>2)=1-[P(r=0)+P(r=1)+P(r=2)]$

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

$$=1-\frac{5}{e^2}=0.323.$$ **Ans.**

(iii) P (none suffers a bad reaction) $=P(r=0)=e^{-m}=1/e^2=0.135.$ **Ans.**

(iv) P (more than 1 suffers a bad reaction)

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

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

$$=1-\frac{3}{e^2}=0.594.$$ **Ans.**

Example 5.29 : A car-hire firm two cars, which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused. ***[RGPV Feb. 2006]***

Solution. Given mean $m=1.5$,

Poisson distribution $P(r)=\dfrac{e^{-m}m^r}{r!}$, $r=0,1,2,3,\ldots\ldots\ldots\ldots$

(i) The proportion of days when neither car is used

i.e., $P(r=0)=\dfrac{e^{-m}\ m^0}{0!}=$

$e^{-1.5}=0.2231.$ **Ans.**

(ii) Since total cars = 2

$\therefore$ demand is refused when $r\geq 3$.

Hence, the proportion of days on which some demand is refused

$$=P(r\geq 3)=1-[P(r=0)+P(r=1)+P(r=2)]$$

$$=1-\left[e^{-m}\frac{m^0}{0!}+e^{-m}\frac{m^1}{1!}+e^{-m}\frac{m^2}{2!}\right]$$

$$=1-e^{-1.5}\left[1+(1.5)+\frac{(1.5)^2}{2}\right]$$

$$=1-0.8087\quad=0.1913.$$ **A**

Example 5.30 : ***A telephone switch handles 600 calls on the average during a rush hour. The board can make a maximum 20 connections per minute. Use Poise n distribution to estimate the probability that the board will be over-taxed during any given minute.***

Solution. Given that mean $m =$ number of calls per minute

or $$m = \frac{600}{60} = 10.$$

$\therefore$ Poisson distribution $P(r) = e^{-m}\frac{m^r}{r!}$, $\quad r = 0, 1, 2, \ldots\ldots\ldots 600.$

$\therefore$ P (using 0 to 20 callas per minute) $= P(r \le 20)$

$$= P(r=0) + P(r=1) + \ldots\ldots + P(r=20)$$

$$= \sum_{r=0}^{20} e^{-m}\frac{m^r}{r!} = e^{-10}\sum_{r=0}^{20}\frac{m^r}{r!}, \qquad \ldots(1)$$

Thus P (the board will be over-taxed during any given minute)

$= P$ (when the calls are more than 20)

$= P(r > 20) = 1 - P(r \le 20)$

$$= 1 - e^{-10}\sum_{r=0}^{20}\frac{m^r}{r!}.$$ **Ans.**

Example 5.31 : ***If 3% of the electric bulbs manufactured by a company are defective, find the probability that in a sample of 100 bulbs exactly five bulbs are defective.***

Solution. Given that $\quad p = \frac{3}{100} = 0.03, \quad n = 100$

$\therefore$ mean $\quad m = np = 100 \times 0.03$

$\Rightarrow \quad m = 3.$

By Poisson distribution $P(r) = e^{-m}\frac{m^r}{r!}$, $\quad r = 0, 1, 2, \ldots\ldots\ldots 100.$

$\therefore$ P (exactly five bulbs are defective) $= P(r = 5)$

$$= e^{-3}\frac{(3)^5}{5!} = \frac{0.04979 \times 243}{120} = 0.1008.$$ **Ans.**

cample 5.32 : ***A manufacture knows that the condensers he makes contain on an average 1% of defective He packs them in boxes of 100. What is the probability that a box picked at random will contain 4 or more faulty condensers ?***

lution. Given that $p = \frac{1}{100} = .01, \quad n = 100$. Mean $= m = np = 100 \times 0.01 = 1.$

By Poisson distribution :

$$P(r)=\frac{e^{-m}(m)^r}{r!}, \qquad r=0,1,2.........100.$$

$\therefore$ P (4 or more faulty condensers)

$$=P(r\geq 4)=P(4)+P(5)+....+P(100)$$

$$=1-[P(0)+P(1)+P(2)+P(3)]$$

$$=1-\left[\frac{e^{-1}}{0!}+\frac{e^{-1}}{1!}+\frac{e^{-1}}{2!}+\frac{e^{-1}}{3!}\right]=1-e^{-1}\left[1+1+\frac{1}{2}+\frac{1}{6}\right]$$

$$=1-\frac{8}{3e}=1-0.981=0.019.$$

Ans.

xample 5.33 : A Skilled typist on routine work kept a record of mistakes made per day during 300 working days

Mistake per day :	***0***	***1***	***2***	***3***	***4***	***5***	***6***
Number of days :	***143***	***90***	***42***	***12***	***9***	***3***	***1***

Fit a Poisson distribution to the above date and calculate expected (or theoretical) frequencies

olution. Mean $=m=\dfrac{143\times0+90\times1+42\times2+12\times3+9\times4+3\times5+1\times6}{300}$ $[\because N=300]$

or $\quad m=\dfrac{0+90+84+36+15+6}{300}=\dfrac{267}{300}=0.89 \quad m=0.89.$

By Poisson distribution :

$$P(r)=\frac{e^{-m}m^r}{r!}, \qquad r=0,1,2,.........6.$$

$\therefore$ The expected (or theoretical) frequency for r success

$$f(r)=N.P(r)=N.\frac{e^{-m}m^r}{r!},\ r=0,1,2,......6.$$

for $r=0$, $\quad f(0)=NP(0)=\dfrac{300e^{-0.89}(0.89)^0}{0!}=300\times0.411=123.3\approx123$

for $r=1$, $\quad f(1)=NP(1)=\dfrac{300e^{-0.89}(0.89)}{1!}=300\times0.365=109.5\approx110$

for $r=2$, $\quad f(2)=NP(2)=\dfrac{300e^{-0.89}(0.89)^2}{2!}=300\times0.163=48.9\approx49$

for $r=3$, $\quad f(3)=NP(3)=\dfrac{300e^{-0.89}(0.89)^3}{3!}=300\times0.048=14.4\approx14$

for $r=4$, $\quad f(4)=NP(4)=\dfrac{300e^{-0.89}(0.89)^4}{4!}=300\times0.011=3.3\approx3$

for $r=5$, $$f(5)=NP(5)=\frac{300e^{-0.89}(0.89)^5}{5!}=300\times 0.002=0.6\approx 1$$

for $r=6$, $$f(6)=NP(6)=\frac{300e^{-0.89}(0.89)^6}{6!}=300\times 0.0003=0.09\approx 0.$$

Thus expected (or theoretical) frequencies are :

r :	0	1	2	3	4	5	6
f_e :	123	110	49	14	3	1	0.

Example 5.34 : ***The frequency of accidents per shift in a factory is given in the table below***

Accident per shift (x)	0	1	2	3	4	*Total*
Frequency (f)	192	100	24	3	1	320.

Find the corresponding Poisson distribution and compare with actual observation.

Solution. Mean $=m\frac{\Sigma fx}{\Sigma f}=\frac{0\times 192+1\times 100+2\times 24+3\times 3+4\times 1}{192+100+24+3+1}$

$$=\frac{161}{320}=0.503.$$

By Poisson distribution, $P(x=r)=\frac{e^{-m}m^r}{r!}$, $r=0,1,2,3,4.$

$\therefore$ The corresponding probabilities for r success

$$P(r)=e^{-0.503}\frac{(0.503)^r}{r!},\quad r=0,1,2,3,4.$$

i.e., Probabilities are : 0.605, 0.304, 0.076, 0.0128, 0.0016.

Now the corresponding expected frequency for r success

$$f(r)=N.P(r)=320\times e^{-0.503}\frac{(0.503)^r}{r!},\quad r=0,1,2,3,4.$$

i.e., Corresponding frequencies are :

193.6, 97.3, 24.5, 4.1, 0.5.

Example 5.35 : ***Fit a Poisson's distribution to the following calculate theoretical frequencies.***

7.3 Regression

A regression model is a mathematical equation that describes the relationship between two or more variables. It is also known as regression equation.

Let us consider the example of relationship between food expenditure and income. But food expenditure can be affected by many other variables like tastes and preferences of household members or the size of the household etc. These variables are called independent or explanatory variables as they vary independently and explain the changes in food expenditure among different households, while the food expenditure is called the dependent variable because it depends on the above given independent variables.

Hence in regression analysis the *dependent variable* is one whose value is influenced or to be predicted. It is also known as regressed or explained variable while the variable which influences the values and is used for predicting the values of dependent variable is called as independent variable or regressor or predictor or explanatory variable.

Simple Regression:

If we study the effect of a single independent variable on a dependent variable, it is called simple regression and such a model is known as simple regression model.

Multiple Regression

Studying the effect of two or more independent variables on a dependent variable is known as multiple regression and such a model is known as multiple regression model.

7.3.1 Linear Regression

A regression equation, when plotted, may assume one of many possible shapes known as curve of regression. If this curve of regression is a straight line then it is said to be ***line of Regression*** and such a regression is said to be linear regression. If the curve of regression is not a straight line then the regression is known as curvilinear regression or non linear regression.

Definition

A simple regression model that gives a straight line relationship between two variables is said to be **linear regression model.**

We always have two lines of regression in a simple linear regression model. Let the equation of the linear relationship between the two variables x and y be of the form y = a + bx, where y is treated as dependent variable and x is treated as independent variable. On treating these other way round i.e., considering x as dependent variable and y as independent variable we can have the linear equation of the form x = c + dy; thus we have two lines of regression. The lines of regression give the best estimate to the values of one variable for any specific value of the other variable.

Remark 9. As the line of regression is the line of best fit, it is obtained by principles of least squares.

7.3.2 Lines of Regression

1. Equation of line of regression of Y on X

If we choose the straight line in a linear regression model such that the sum of squares of deviations parallel to the axis of *y* is minimized, it is called the line of regression of Y on X. It gives the best estimates of *Y* for any given value of *X*.

Derivation

Let y = a + bx be the line of regression of *Y* on *X* for the given data (x_i, y_i), $i = 1, 2, \ldots n$. .....(1)

Then as discussed in section 3.6.1 of chapter 3, for

$$y = a + bx$$

the normal equations are

$$\sum_{i=1}^{n} y = \sum_{i=1}^{n} a + \sum_{i=1}^{n} bx \Rightarrow \sum_{i=1}^{n} y = na + b\sum_{i=1}^{n} x \qquad \ldots(2)$$

and

$$\sum_{i=1}^{n} xy = \sum_{i=1}^{n} ax + \sum_{i=1}^{n} bx^2 \Rightarrow \sum_{i=1}^{n} xy = a\sum_{i=1}^{n} x + b\sum_{i=1}^{n} x^2 \qquad \ldots(3)$$

(Here the summation $\left(\sum_{i=1}^{n}\right)$ corresponds to values (x_i, y_i) $i = 1, 2, \ldots, n$).

Dividing equation (2) by n we have :-

$$\frac{\Sigma y}{n} = a + b\frac{\Sigma x}{n} \Rightarrow \bar{y} = a + b\,\bar{x} \qquad \ldots(4)$$

as,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\Sigma x}{n}$$

and

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\Sigma y}{n}$$

Again dividing equation (3) by n we have :-

$$\frac{\Sigma xy}{n} = a\frac{\Sigma x}{n} + b\frac{\Sigma x^2}{n} \Rightarrow \frac{\Sigma xy}{n} = a\bar{x} + b\frac{\Sigma x^2}{n} \qquad \ldots(5)$$

Now as studied in sec 7.2.1

$$Cov(x, y) = \mu_{11} = \frac{\Sigma xy}{n} - \bar{x}\,\bar{y} \Rightarrow \frac{\Sigma xy}{n} = \mu_{11} + \bar{x}\,\bar{y} \qquad \ldots(6)$$

Also

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - \bar{x}^2 \Rightarrow \frac{\Sigma x^2}{n} = \sigma_x^2 + \bar{x}^2 \qquad \ldots(7)$$

Now equations (5) to (7) imply that

$$\mu_{11} + \bar{x}\,\bar{y} = a\bar{x} + b\left(\sigma_x^2 + \bar{x}^2\right)$$

$$\Rightarrow \quad \mu_{11} + \bar{x}\,\bar{y} = a\bar{x} + b\sigma_x^2 + b\bar{x}^2 = \bar{x}(a + b\bar{x}) + b\sigma_x^2$$

$$= \bar{x}\,\bar{y} + b\sigma_x^2$$

(using equation (4))

$$\Rightarrow \quad \mu_{11} = b\sigma_x^2$$

$$\Rightarrow \qquad b = \frac{\mu_{11}}{\sigma_x^2} = \frac{r\sigma_x\,\sigma_y}{\sigma_x^2} = \frac{r\sigma_y}{\sigma_x} \qquad \left(\because r = \frac{\mu_{11}}{\sigma_x\,\sigma_y}\right)$$

From equation (1) and equation (4) it is quite clear that the required line passes through $(\bar{x}, \bar{y})$. Hence the equation of line passing through point $(\bar{x}, \bar{y})$ and having slope $b_{yx} = r\sigma_y/\sigma_x$ is :-

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

$$\Rightarrow \qquad y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \text{.....(8)}$$

which is the required line of regression of Y on X.

2. Equation of line of regression of X on Y

In the linear reression model if the straight line is so chosen that the sum of squares of deviations parallel to axis of x are minimized, then it is called as line of regression of Y on X. Here x is treated as dependent variable and y is treated as independent variable.

It gives the best estimates of x for any given value of y.

It can be derived in the same manner as done in part (1) above, by interchanging the role of x and y. Its equation will be :-

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \text{.....(25)}$$

In case of perfect correlation (i.e. $r = \pm 1$) the equation of line of regression of y on x is

$$y - \bar{y} = \frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \text{(using 8)}$$

and equation of line of regression of x on y is

$$x - \bar{x} = \frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \text{(using (9))}$$

both of which are similar.

Hence in general, we always have two lines of regression except in the case of perfect correlation ($r = \pm 1$).

Remark 10. The line of regression of y on x as well as that of x on y both pass through $(\bar{x}, \bar{y})$. Hence $(\bar{x}, \bar{y})$ is the point of intersection of two lines of regression.

Remark 11. The equations for both the lines of regression are not reversible or interchangeable as basis and assumptions for deriving these equations are quite different.

Remark 12. If we have to predict the values of y for a given value of x then line of egression of y on x must be used, as in this case the predicted values will have minimum possible error (as obtained by principle of least squares).

Properties of Regression Coefficient

(i) We know that $b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$ is regression coefficient of y on x and $b_{xy} = r\dfrac{\sigma_x}{\sigma_y}$ is regression coefficient of x on y, hence $b_{yx} \cdot b_{xy} = r^2$.

$\Rightarrow$ $r = \pm\sqrt{b_{yx} \cdot b_{xy}}$ and the sign of r is same as that of the two regression coefficients.

(ii) We know that $r^2 \leq 1 \Rightarrow b_{yx}\ b_{xy} \leq 1$

$$\Rightarrow \qquad b_{xy} \leq \frac{1}{b_{yx}}$$

Now if $b_{yx} > 1 \Rightarrow b_{xy} < 1$. Hence if one of the regression coefficient is greater than unity, the other must be less than unity.

(iii) Arithmetic mean of regression coefficient is greater than the correlation coefficient (r), provided r > 0.

Arithmetic mean of regression coefficients $= \dfrac{1}{2}\left(b_{yx} + b_{xy}\right) = \dfrac{1}{2}\left(r\dfrac{\sigma_y}{\sigma_x} + r\dfrac{\sigma_x}{\sigma_y}\right)$

Now $\qquad \left(\sigma_y - \sigma_x\right)^2 = 0 \Rightarrow \sigma_y^2 + \sigma_x^2 - 2\sigma_x\sigma_y \geq 0 \Rightarrow \dfrac{\sigma_x}{\sigma_y} + \dfrac{\sigma_y}{\sigma_x} \geq 2$

$$\Rightarrow \qquad r\left(\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y}\right) \geq 2r \qquad [\because r > 0]$$

$$\Rightarrow \qquad \frac{1}{2}r\left(\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y}\right) \geq r$$

Hence Proved

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

7.3.3 Angle Between two Lines of Regression

Equation of line of regression of y on x is

$y - \bar{y} = r\dfrac{\sigma_y}{\sigma_x}(x - \bar{x})$ whose slope is $b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$

uation of line of regression of x on y is :-

$x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \Rightarrow y-\bar{y}=\frac{\sigma_y}{r\sigma_x}(x-\bar{x})$ whose slope is $\frac{1}{b_{xy}}=\frac{\sigma_y}{r\sigma_x}$

θ is the angle between the two lines of regression then

$$\tan\theta=\frac{\frac{\sigma_y}{r\sigma_x}-\frac{r\sigma_y}{\sigma_x}}{1+\frac{r\sigma_y}{\sigma_x}\frac{\sigma_y}{r\sigma_x}}=\frac{\left(1-r^2\right)\left(\sigma_y/\sigma_x\right)}{\sigma_x^2+\sigma_y^2}\times\frac{\sigma_x^2}{r}$$

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

$$\Rightarrow \qquad \theta=\tan^{-1}\left\{\frac{1-r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right\}$$

As $r^2\le 1 \Rightarrow 1-r^2\ge 0 \Rightarrow 0\le\theta<\pi/2$

Hence the acute angle (θ_1) between the two lines is

$$\tan\theta_1=\left(\frac{1-r^2}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

Conversely $r^2\le 1\Rightarrow r^2-1\le 0 \Rightarrow \pi/2<Q\le\pi$, hence the obtuse angle (θ_2) between the two lines is

$$\tan\theta_2=\left(\frac{r^2-1}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

Hence we have the following possible cases:

(i) If $r=0\Rightarrow\tan\theta=\infty\Rightarrow\theta=\frac{\pi}{2}$ **[Raj. IV Sem CP-2003]**

Hence if the two variables are uncorrelated, the lines of regression become perpendicular to each other. Here as r=0 the lines of regression are

$$y-\bar{y}=0\Rightarrow y=\bar{y} \text{ and } x-\bar{x}=0\Rightarrow x=\bar{x}$$

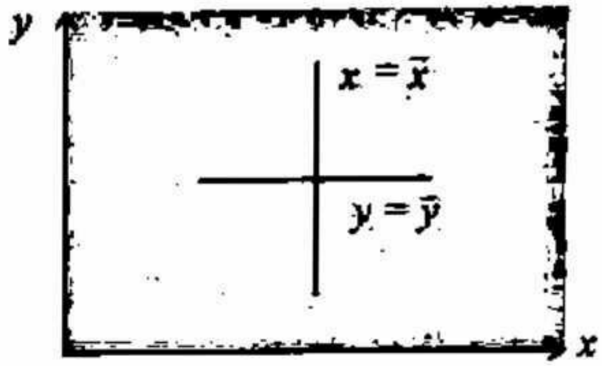


Figure 7.5

(ii) If $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$ or π. **(Raj. IV Sem CP-2003)**

This means that either the two lines are parallel to each other or the two lines coincide. But we know that both the lines intersect at the point $(\bar{x}, \bar{y})$, hence they cannot be parallel and must be coincident. Therefore in case of perfect correlation the two lines of regression are coincident.

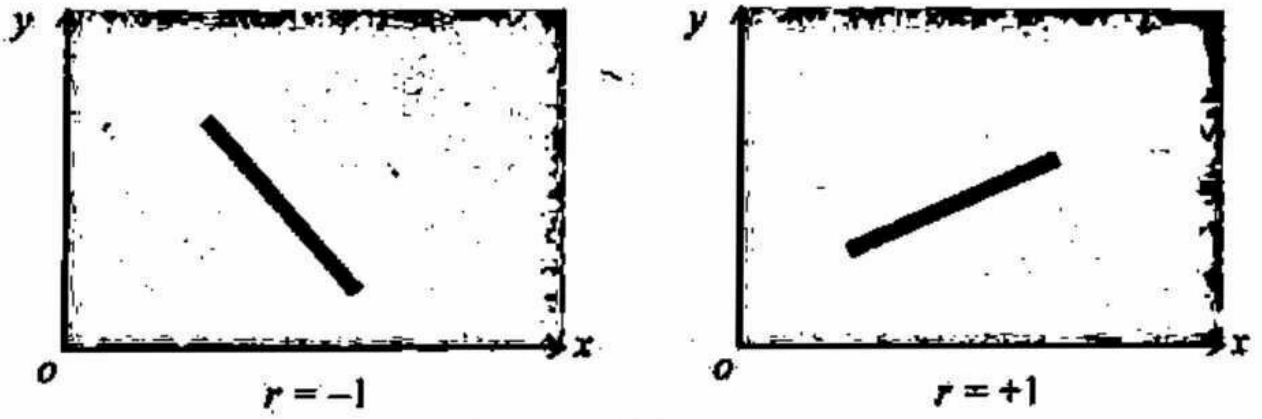
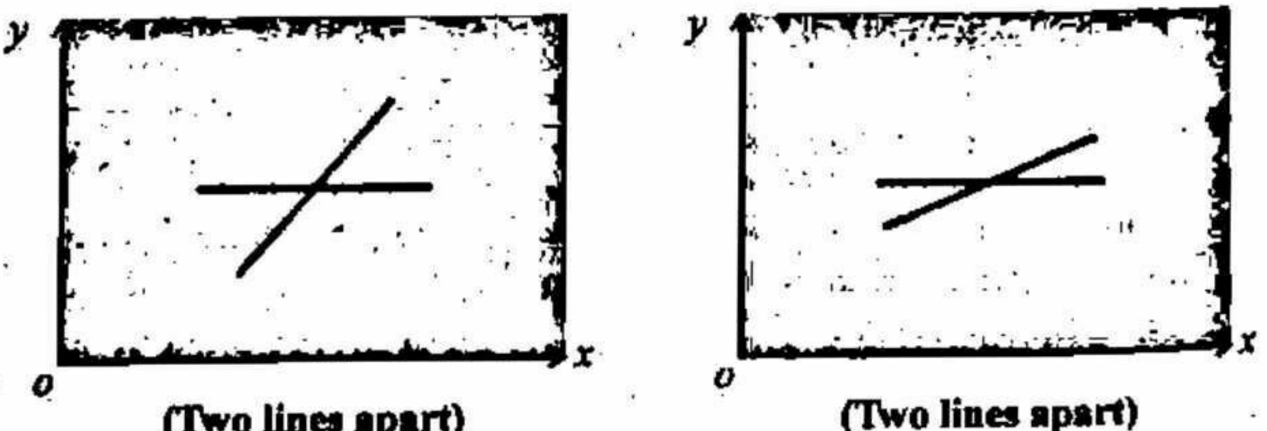


Figure 7.6

Remark 13. $r = 0 \Rightarrow \theta = \pi/2$ *and* $r = \pm 1 \Rightarrow \theta = 0$.

Hence we can conclude that for higher degree of correlation between the variables the angle between the lines is smaller, i.e., the two lines of regression are closer to each other and similarly by same reasoning we can say that larger angle between them indicates a poor degree of correlation between the variables. Thus by plotting the lines of regression on graph paper we can have a rough idea about the degree of correlation between the two variables.



(Two lines apart)
Low degree of correlation

(Two lines apart)
High degree of correlation

Figure 7.7

Ex.7 Calculate the coefficient of correlation and obtain the line 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

[Raj. IV Sem CP-2004]

Sol. As done in Ex. 4 of this chapter, we solve to get

$r_{xy} = r_{uv} = 0.95$

As $u = x - 5$ and $v = y - 12$

$\Rightarrow \bar{u} = \bar{x} - 5$ and $\bar{v} = \bar{y} - 12$

$$\sigma_u^2 = E\left\{(u-\bar{u})^2\right\} = E\left[\left[(x-5)-(\bar{x}-5)\right]^2\right] = E\left[(x-\bar{x})^2\right] = \sigma_x^2$$

and
$$\sigma_v^2 = E\left[(v-\bar{v})^2\right] = E\left[\left[(y-12)-(\bar{y}-12)\right]^2\right] = E\left[(y-\bar{y})^2\right] = \sigma_y^2$$

Line of Regression of y on x is :-

$$y - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$\Rightarrow \quad y - (\bar{v}+12) = r_{uv}\frac{\sigma_v}{\sigma_u}\left[x-(\bar{u}+5)\right]$$

$$\Rightarrow \quad y-(0+12) = 0.95 \times \frac{\sqrt{60/9}}{\sqrt{60/9}}\left[x-(0+5)\right] \quad \text{(refer Ex. 4 of this chapter)}$$

$$\Rightarrow \quad y - 12 = 0.95\,(x-5) \quad \Rightarrow \quad y = 0.95x + 7.25$$

Line of Regression of x on y is :-

$$x - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

$$\Rightarrow \quad x - (\bar{u}+5) = r_{uv}\left[y-(\bar{v}+12)\right] \quad \left(\sigma_x = \sigma_y \ as\ \sigma_u = \sigma_v\right)$$

$$\Rightarrow \quad x - 5 = 0.95\,(y-12)$$

$$\Rightarrow \quad x = 0.95\,y - 6.4$$

Ex.8 In a partially destroyed laboratory on record of an analysis of correlation data, the following results only are legible,

Var x = 9, Regression equations : $8x - 10y + 66 = 0$, $40x - 18y = 214$

Find

(i) The mean values of x and y.

(ii) The standard deviation of y.

(iii) The coefficient of correlation between x and y. [Raj. IV Sem CP-2005]

Sol. (i) We know that the mean value is the common point of intersection of the two lines of regression. Given regression equations are

$$8x - 10y + 66 = 0$$

$$40x - 18y = 214$$

Solving the above two equations we get $x = 13$ and $y = 17$.

Hence the mean values are $\bar{x} = 13, \bar{y} = 17$.

(ii) & (iii)First regression equation $\Rightarrow y = \frac{8}{10}x + \frac{66}{10}$

which can be treated as line of regression of y on x and second regression equation

$$\Rightarrow \qquad x = \frac{18}{40}y + \frac{214}{40}$$

which can be treated as line of regression of x on y

$$\Rightarrow \qquad b_{yx} = \frac{8}{10} \text{ and } b_{xy} = \frac{18}{40}$$

As $$r^2 = b_{yx} \times b_{xy} = \frac{8}{10} \times \frac{18}{40} = \frac{9}{25} = 0.36 \quad \Rightarrow r = \pm 0.6$$

As both regression coefficients b_{yx} and b_{xy} are positive hence the correlation coefficient should also be positive and r = 0.6.

Moreover $$b_{yx} = \frac{r\sigma_y}{\sigma_x} = \frac{8}{10}$$

Here $r = 0.6$, $\sigma_x = \sqrt{9} = 3$ *(given)*

$$\Rightarrow \qquad 0.6 \times \frac{\sigma_y}{3} = \frac{8}{10} \Rightarrow \sigma_y = \frac{4}{5} \times \frac{1}{0.2} = 4$$

i.e., $$\sigma_y = 4,$$

Remark 14. If we take the first regrassion equation as line of regression of x on y i.e., $x = \frac{10}{8}y - \frac{66}{8}$ and the second regression equation as line of regression of y on x

i.e., $y = \frac{40}{18}x - \frac{214}{18}$ then $r^2 = \frac{10}{8} \times \frac{40}{18} = 2.778$

$\Rightarrow r = 1.6.$

Ex.9 Calculate the coefficient of correlation between x and y using the following data.

x	-10	-5	0	5	10
y	5	9	7	11	13

Sol. $r=\dfrac{Cov(x,y)}{\sigma_x\sigma_y}; Cov(x,y)=\dfrac{\Sigma xy}{n}-\bar{x}\,\bar{y}, \sigma_x=\sqrt{\dfrac{\Sigma x^2}{n}-(\bar{x})^2}, \sigma_y=\sqrt{\dfrac{\Sigma y^2}{n}-(\bar{y})^2}$

We construct the following table :-

	x	y	x^2	y^2	xy
	-10	5	100	25	-50
	-5	9	25	81	-45
	0	7	0	49	0
	5	11	25	121	55
	10	13	100	169	130
Total	0	45	250	445	90

Hence $\bar{x}=\dfrac{\Sigma x}{5}=0,\quad \bar{y}=\dfrac{\Sigma y}{5}=9$

$$Cov(x,y)=\frac{\Sigma xy}{n}-\bar{x}\,\bar{y}=\frac{90}{5}-0=18$$

$$\sigma_x=\sqrt{\frac{\Sigma x^2}{n}-(\bar{x})^2}=\sqrt{\frac{250}{5}}=\sqrt{50}=7.0711$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{n} - (\bar{y})^2} = \sqrt{\frac{445}{5} - 81} = \sqrt{8} = 2.8284$$

Hence $$r = \frac{18}{7.0711 \times 2.8284} = 0.90.$$

Ex.10 Calculate *Cov* (*x*,*y*) when $\Sigma x = 50$, $\Sigma y = -30$, $\Sigma xy = -115$, $n = 10$.

Sol. We know that

$$Cov(x,y) = \frac{\Sigma xy}{n} - \bar{x}\,\bar{y}$$

$$= \frac{-115}{10} - \left(\frac{50}{10} \times \frac{-30}{10}\right) \qquad \left(\because \bar{x} = \frac{\Sigma x}{n}, \bar{y} = \frac{\Sigma y}{n}\right)$$

$$= \frac{-115}{10} + 15 = 3.5$$

Ex.11 For a bivariate distribution n = 18, $\Sigma x^2 = 60$, $\Sigma y^2 = 96$, $\Sigma x = 12$, $\Sigma y = 18$, $\Sigma xy = 48$. Find the equations of the lines of regression and r. **[Raj. IV Sem CP-2006]**

Sol. $$\bar{x} = \frac{\Sigma x}{n} = \frac{12}{18} = 0.667, \bar{y} = \frac{\Sigma y}{n} = \frac{18}{18} = 1$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2 = \frac{60}{18} - (0.667)^2 = 2.8884$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - (\bar{y})^2 = \frac{96}{18} - 1 = 4.3333$$

$$Cov(x,y) = \frac{\Sigma xy}{n} - \bar{x}\,\bar{y} = \frac{48}{18} - (0.667)(1)$$

$$= 2.6667 - 0.667 = 1.9997.$$

Hence line of regression of y on x is :-

$$(y - \bar{y}) = \frac{Cov(x,y)}{\sigma_x^2}(x - \bar{x}) \Rightarrow (y-1) = \frac{1.9997}{2.8884}(x - 0.667)$$

$$\Rightarrow \quad y - 1 = 0.69232\,(x - 0.667)$$

$$\Rightarrow 0.69232x - y + 0.538 = 0$$

$$\Rightarrow \quad y = 0.692x + 0.538.$$

Con

Similarly line of regression of x on y is :-

$$(x-\bar{x}) = \frac{Cov(x,y)}{\sigma_y^2}(y-\bar{y}) \Rightarrow (x-0.667) = \frac{1.9997}{4.3333}(y-1)$$

$$\Rightarrow \quad x - 0.667 = 0.4615(y-1)$$

$$\Rightarrow \quad x = 0.4615y + 0.2055$$

Also coefficient of correlation $r^2 = 0.4615 \times 0.692 = 0.3194$

$$\Rightarrow \quad r = 0.565 \cong 0.57.$$

.12 Two random variables have the least square regression lines with equations :-

$$3x + 2y - 26 = 0 \quad \text{and} \quad 6x + y - 31 = 0.$$

Find the mean values and coefficient of correlation between x and y. ii) if variance of y is 4 find S.D of X

ol. The given regression equations are :-

$$3x + 2y - 26 = 0 \qquad \text{.....(1)}$$

and

$$6x + y - 31 = 0 \qquad \text{.....(2)}$$

Let equation (1) be line of regression of y on x

$$\Rightarrow \quad y = -\frac{3}{2}x + 13$$

Let equation (2) be line of regression of x on y

$$\Rightarrow \quad x = -\frac{1}{6}y + \frac{31}{6}.$$

Hence regression coefficients are $b_{yx} = \frac{-3}{2}$ and $b_{xy} = \frac{-1}{6}$

$$\Rightarrow \quad r^2 = b_{yx} \times b_{xy} = \frac{3}{2} \times \frac{1}{6} = \frac{1}{4} \Rightarrow r = \pm\frac{1}{2}$$

But as b_{yx} and b_{xy} are negative in sign hence

$$r = -\frac{1}{2} = -0.5$$

Again solving (1) and (2) as simultaneous equations as :-

[equation (2)] × 2 – equation (1)

$$\Rightarrow \quad 9x - 36 = 0 \Rightarrow x = 4$$

Substituting it in (1) we get :-

$$2y = 26 - 3 \times 4 = 26 - 12 = 14 \quad \Rightarrow y = 7$$

Now as the mean values are points of intersection of regression equations (1) and (2), hence we get the mean values as

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

Remark 15. If we consider equation (1) to be line of regression of x on y i.e., $x = -\frac{2}{3}y + \frac{26}{3}$ and equation (2) be line of regression of y on x i.e., $y = -6x+31$ then $r^2 = (-6)(-2/3) = 4 > 1$, hence not possible.

The ranking of students in two subjects A and B are as follows :-

A	3	5	8	4	7	10	2	1	6	9
B	6	4	9	8	1	2	3	10	5	7

What is the coefficient of rank correlation ?

We form the following table :-