7. MATHEMATICAL EXPECTATION GENERATING FUNCTIONS AND LAW OF LARGE NUMBERS

7.1 Mathematical Expectation

The mathematical expectation of a discrete random variable x having values $x_1, x_2, \ldots\ldots x_n$ with respective probabilities $P(x_1)$, $P(x_2) \ldots\ldots P(x_n)$ is defined by

$$E(x)= \sum^{n} x_i(Px_i) \text{ where } \sum^{n} P(x_i)=1, \qquad \ldots\ldots(7.1\ a)$$

provided the series is absolutely convergent. For example the $E(x!)$ does not exist for the following probability function of x,

$$P(x) = \frac{e^{-1}}{x!} \qquad x=0, 1, 2, \ldots\ldots \propto$$

$$\text{We know, } E(x!) = \sum^{\propto} x!\, P(x) = \sum^{\propto} x! \frac{e^{-1}}{x!} = \sum^{\propto} e^{-1}$$

which is a divergent series. Hence the expected value is not defined.

If x is a continuous random variable with p, d, f, f(x)

$$\text{then, } E(x) = \int_{-\propto}^{\propto} xf(x)dx, \qquad \ldots\ldots (7.1.b)$$

provided the integral is absolutely convergent.

Remarks :

1) $E(a)=a$, where a is a constant.
2) $E(ax)=aE(x)$
3) The mathematical expectation of $\psi(x)$, a function of the variable x is given by

$E(\psi(x))=\sum^{n}\psi(x)\ P(x)$; if x is a discrete variable, and

$E[\psi(x)]=\int_{-\infty}^{\infty}\psi(x)\ f(x)\ dx$; if x is a continuous variable.

7.2 Moments

If $\psi(x)=x^r$, then $E(x^r)=\sum^{n} x_i^r\ P(x_i)$ .... (7.2.a)

for discrete random variable x_i and

$E(x^r)=\int_{-\infty}^{\infty} x^r\ f(x)\ dx$ ....(7.2.b)

for continuous random variable x, $-\infty \leq x \leq \infty$.

$E(x^r)$ in both the case is called the rth raw moment of the distribution usually denoted by μ'_r. Thus ,

$\mu'r = E(x^r)$, in particular,

$\mu'_1 = E(x) = \mu$, the mean of the distribution.

$\mu'_2 = E(x^2)$ and $\mu_2 = \mu'_2-\mu'_1 = (E(x^2) - [E(x)]^2$

$=$ var (x) $= \sigma^2$, the variance of the distribution.

f $\psi(x) = x-\mu^r$, then $E(x-\mu)^r =)\sum^{n}(x_i - \mu)^r\ P(x_i)$, for discrete random variable x_i\

and $E(x-\mu)^r = \int_{-\infty}^{\infty}(x-\mu)^r\ f(x)\ dx$, $-\infty \leq x \leq \infty$, for continuous random variable x.

$E(x-\mu)^r$ in both the cases is called corrected rth moment or the rth moment about mean and usually denoted by μ_r.

In particular, if r = 1, $\mu_1 = E(x-\mu) = 0$.

and r = 2, $\mu_2 = E(x-\mu)^2 = E[x - E(x)]^2 = var(x)= \sigma^2$

Example 7.1 Find the expected value of the number of points that will be obtained in a single toss of a fair die.

Solution : Here the variate x is the number of points on a die. Hence the possible values of x are 1, 2, 3, 4, 5 and 6, and each having the probability $\frac{1}{6}$.

Therefore, $E(x) = \sum^{n} x_i p(x_i) = \frac{1}{6}(1+2+3+4+5+6) = \frac{6 \times 7}{6 \times 2} = \frac{7}{2} = 3.5$

Example 7.2 Find the expectation of x whose p. d. f. is $f(x) = 3x^2;\ 0 \le x \le 1$.

Solution : We know, $E(x) = \int_0^1 x f(x)\,dx = \int_0^1 x.3x^2 dx$

$$= \int_0^1 3x^3 dx = \left.\frac{3x^4}{4}\right]_0^1 = \frac{3}{4}.$$

Theorem 7.1 Additive Law of Expectation : The expectation of the sum of two random variables is equal to the sum of their expectations. Symbolically, if x and y are two random variables, then,

$E(x+y) = E(x) + E(y)$ ....(7.3)

Proof : (For discrete variable)

Let P_{ij} be the probability that x assumes the value x_i (i=1,2,....m). and y assumes the value y_j (j=1, 2,.....n). Then

$$E(x+y) = \sum^{m}\sum^{n} (x_i + y_j)\, p_{ij}$$

$$= \sum_i\sum_j x_i p_{ij} + \sum_i\sum_j y_j p_{ij} = \sum_i x_i \sum_j p_{ij} + \sum_j y_j \sum_i p_{ij}$$

$$\left[\text{Since } \sum_i p_{ij} = p_j \text{ and } \sum_j p_{ij} = p_i\right]$$

$$= \sum x_i p_i + \sum y_i p_j = E(x) + E(y).$$

(For Continuous variable)

Let f(xy) be the joint p. d. f. of the random variables x and y, then by definition,

$$E(x+y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f(xy)\, dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(xy)\, dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(xy)\, dxdy.$$

$$= \int_{-\infty}^{\infty} x f(x)\, dx + \int_{-\infty}^{\infty} y f(y)\, dy = E(x) + E(y).$$

Remarks :

1) The above theorem can be generalised for several random variables i, e, if x, y, z....etc. are several random variable then

   $E(x+y+z+......) = E(x) + (E(y) + E(z) + ...$

2) $E(ax+by) = aE(x) + bE(y)$, where a and b are constants.

3) $E[\psi_1(x) + \psi_2(y)] = E(\psi_1(x)] + E[\psi_2(y)]$ where $\psi_1(x)$ and $\psi_2(y)$ are two functions of random variables x and y respectively.

Example 7.3 Find the expected value of the number of points that wil be obtained in a single toss of n fair dice.

Solution : Let x_i be the number of points obtained from the ith die (i=1, 2, ... n) and let $S = x_1 + x_2 + \ldots \ldots + x_n$.

By definition $E(s) = E(x_1 + x_2 + \ldots \ldots + x_n) = E(x_1) + E(x_2) + \ldots \ldots + E(x_n)$.

But for every single die $E(x_i) = \frac{7}{2}$. (i=1, 2... ... n) (vide Example 7.1)

Therefore, $E(s) = \frac{7n}{2} = 3.5n$.

Theorem 7.2 Multiplicative Law of Expectation : The expectation of the product of two independent random variables is equal to the product of their expectations. Symbolically if x and y are two independent random variables, then

$E(xy) = E(x) E(y)$ .... (7.4)

Proof : (For discrete variables)

Let the probability of the discrete random variable x assuming the values x_i, (i=1, 2, ...m) be p_i and that of y assuming the values y_j (j=1, 2,......n) be p_j. Since x and y are independent variables, the probability that the product will assume any value x_iy_j is p_ip_j.

Hence, $E(xy) \sum_{i=1}^{m} \sum_{j=1}^{n} x_iy_jp_ip_j = \sum_{i=1}^{m} x_ip_i \sum_{j=1}^{n} y_jp_i = E(x).\ E(y).$

(For continuous variables)

Let f(x, y) be the joint p. d. f. of the joint random variables x and y, then by definition,

$$E(xy) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ f(xy)\ dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ f(x)\ f(y)\ dxdy$$

[Since f (xy)=f(x). f(y) for independent random variable x and y.]

$$= \int_{-\infty}^{\infty} x\ f(x)\ dx \int_{-\infty}^{\infty} y\ f(y)\ dy. = E(x)\ E(y).$$

Remarks :

1) The above theorem can be generalised for several independent random variables i, e, if x, y, z ... etc. are several independent random variables then

   E(xyz ....) = E(x) E(y) E(z) ....

2) If $\psi_1(x)$ and $\psi_2(y)$, are two functions of two independent random variables x and y respectively, then

   $E[\psi_1(x)\ \psi_2(y)] = E[\psi_1(x)]\ E[\psi_2(y)]$.

3) E(ax. .by) = abE(x) E(y). For two independent random variables x and y ; a and b are two constants.

Example 7.4 Find the expected value of the product of points that will be obtained in a single throw of n fair dice.

Solution : We obtained in Example 7.1 that the expected value of $x_i = \frac{7}{2}$ where x_i be the number of points obtained on ith die. Therefore, the expected value of product of points obtained is equal to $\left(\frac{7}{2}\right)^n$

7.3 Covariance

If x and y are two random variables, then the covariance between them is defined as

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

$$= E[xy - xE(y) - yE(x) + (E(x)E(y)]$$

$$= E(xy) - E(x)E(y) - E(y)E(x) + E(x)E(y)$$

$$= E(xy) - E(x)E(y) \qquad \ldots\ \ldots\ \ldots(7.5)$$

Remarks :

1) If x and y are independent random variable then $E(xy) = E(x)\,E(y)$ and hence

   $\text{Cov}(xy) = E(xy) - E(x)\,E(y) = 0.$

   Thus the covariance of two independent random variables is equal to zero. The converse is not necessarily true.

2) $\text{Cov}(ax.by) = ab\,\text{Cov}(xy)$, where a and b are two constants.

3) $\text{Cov}(x+a, y+b) = \text{Cov}(x,y)$ where a and b are two constants acting as respective origins.

4) $\text{Cov}\left(\dfrac{x-\mu_x}{\sigma_x}, \dfrac{y-\mu_y}{\sigma_y}\right) = \dfrac{1}{\sigma_x\sigma_y}\text{Cov}(xy).$

   where μ_x, μ_y are the means and σ_x, σ_y are the standard deviations of the random variables x and y respectively.

5) $\text{Cov}(x,x) = V(x).$

Theorem 7.3 Variance of a Linear Combination of Random Variables :

Let $x_1, x_2, \ldots x_n$ be n random variables (not the values of the variable x) then

$$V\left(\sum^{n} a_i x_i\right) = \sum^{n} a_i^2 V(x_i) + 2\sum^{n}\sum^{n}_{i<j} a_i a_j \,\text{Cov}(x_i x_j)$$

Proof : Let $u = a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n$

we know, $E(u) = a_1E(x_1) + a_2E(x_2) + \ldots\ldots + a_nE(x_n)$.

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

Therefore, $V(u) = V \sum^{n} a_i x_i = E[u - E(u)]^2$

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

$+ 2\sum^{n}\sum^{n}_{i \leq j} a_i a_j E[x_i - E(x_i)][x_j - E(x_j)]$

$= a_1^2 V(x_1) + a_2^2 V(x_2) + \ldots\ldots + a_n^2 V(x_n) + 2\sum^{n}\sum^{n}_{i<j} a_i a_j \operatorname{Cov}(x_i x_j)$

$= \sum^{n} a_i^2 V(x_i) + 2\sum^{n}\sum^{n}_{i<j} a_i a_j \operatorname{Cov}(x_i x_j).$

Remarks :

(1) If $a_i=1$; $i=1, 2, \ldots n$; then $\sum^{n} a_i x_i$ reduces to $\sum x_i$ and

$$V\left(\sum^{n} x_i\right) = \sum V(x_i) + 2\sum^{n}\sum^{n}_{i<j} \operatorname{Cov}(x_i x_j).$$

2) If x's are independent pairwise then $\operatorname{Cov}(x_i x_j)=0$

and $V\left(\sum^{n} a_i x_i\right) = \sum^{n} a_i^2 V(x_i)$.

3) $V(x_1 \pm x_2) = V(x_1) + V(x_2) \pm 2 \operatorname{Cov}(x_1 x_2)$.

If x_1 and x_2 are independent,

then $V(x_1 \pm x_2) = V(x_1) + V(x_2)$.

Example 7.5 Suppose x is a random variable for which $E(x) = 10$ and $Var(x) = 25$. Find the positive values of a and b such that $y = ax - b$ has expectation 0 and variance 1.

Solution : Given $E(x) = 10$; $Var(x) = 25$.

According to the problem, we have $E(ax - b) = 0$ and

$V(ax - b) = 1$, or, $a^2 V(x) = 1$.

or, $a^2 . 25 = 1$ $\therefore a = \frac{1}{5}$

Again $E(ax - b) = 0$

or, $aE(x) - b = 0$

or, $aE(x) = b$

$\therefore b = 2$, Since $E(x) = 10$ and $a = \frac{1}{5}$

7.4 Conditional Expectation and Conditional Variance

If x and y are two connected discrete random variables with conditional distribution function P(x/y), then the conditional expectation of the random variable x for given value of y is defined by

$$E(x/y) = \sum^{n} x_i P(x_i/y) \qquad \text{.......(7.6a)}$$

and the conditional variance of x for given y is

$$V(x/y) = E[\{x - E(x/y)\}^2/y]. \qquad \text{.......(7.6b)}$$

Similarly conditional expectation and conditional variance of y for given value of x can also be defined.

Again for continuous random variable x and y

$$E(x/y) = \int_{-\infty}^{\infty} xf(x/y)\,dx \qquad \text{...... ...(7.7a)}$$

$$\text{and } V(x/y) = E[\{x-E(x/y)\ /y] \qquad \text{..........(7.7b)}$$

where f(x/y) is the conditional p. d. f. of the random variable x for given y.

Theorem 7.4 The expected value of x is equal to the expectation of the conditional expectation of x for given y. Symbolically

$$E(x) = E_y[E(x/y)] \qquad \text{.........(7.8)}$$

Proof : (For discrete case)

$$\text{R. H. S.} = E_y[E(x/y)]$$

$$= E_y\,[\sum_{i=1}^{n} x_i\, p(x_i/y)]$$

$$= E_y\left[\sum^{n}\left\{x_1 \frac{(Px_i y)}{P(y)}\right\}\right] = \sum_{y}\left[\sum^{n}\left\{x_i \frac{(Px_i y)}{P(y)}\right\}\right]P(y)$$

$$= \sum_{y}\sum_{i=1}^{n} x_i P(x_i y) = \sum_{i=1}^{n} x_i \sum_{y} P(x_i y)$$

$$= \sum^{n} x_i P(x_i) = E(x) = \text{L. H. S.}$$

Hence the theorem is proved.

(For Continuous Case)

$$\text{R.H.S.} = E_y[E(x/y)] = \int_{-\infty}^{\infty} E(x/y)\, f(y)\, dy.$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x\, f(x/y) dx \right] f(y) dy$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x. \frac{f(xy)}{f(y)}\, dx \right] f(y) dy.$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f(xy)\, dx dy = \int_{-\infty}^{\infty} x\, f(x)\, dx = E(x) = \text{L.H.S.}$$

Hence the theorem is proved.

Theorem 7.5 The variance of x can be regarded as consisting of two parts, the expectation of the conditional variance and the variance of the conditional expectation, symbolically

$$V(x) = E_y[V(x/y)] + V_y[E(x/y)]. \qquad \text{.......(7.9)}$$

Proof : We know, $V(x/y) = E[\{x-E(x/y)\}^2/y]$

$$=E(x^2/y) - [E(x/y)]^2$$

$$\therefore E_y[V(x/y)]=E_y[E(x^2/y)] - E_y[E(x/y)]^2$$

$$=E(x^2) - E_y[E(x/y)]^2 = V(x) + [E(x)]^2 - E_y[(x/y)]^2$$

$$=V(x) + [E_y(x/y)]^2 - E_y[E(x/y)]^2 = V(x) - V_y[E(x/y)]$$

Therefore, $V(x) = E_y[V(x/y)] + V_y[E(x/y)]$.

Example 7.6 Find E(x/y) from the Example 6.4 given in Chapter 6.

Solution : We know, $f(x/y) = (1+y)^2 xe^{-x(1+y)};\ x, y, \geq 0,$

Therefore, $E(x/y) = \int_0^{\infty} x\, f(x/y)\, dx = \int_0^{\infty} x(1+y)^2 xe^{-x(1+y)} dx$

$$= (1+y)^2 \int_0^{\infty} x^2 e^{-x(1+y)}\, dx$$

$$\left[\text{Putting } x(1+y) = z \text{ or, } x = \frac{z}{1+y};\ dx = \frac{dz}{(1+y)}\right]$$

$$= \frac{1}{(1+y)} \int_0^{\infty} z^2 e^{-z} dz$$

$$= \frac{1}{(1+y)} \Gamma 3 \qquad = \frac{2}{(1+y)}.$$

7.5 Moment Generating Function (m. g. f.)

The moment generating function (m. g. f.) of a random variable x about origin is defined as

$$M_0(t) = E(e^{tx}) = \sum_x e^{tx} P(x), \text{ for discrete random variable x and discrete probability distribution.}$$

$$= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx, \text{ for continuous random variable x and continuous probability distribution.} \qquad \text{.....(7.10)}$$

The m. g. f. is a function of the real parameter t and it is being assumed that the right hand side of (7.10) is absolutely convergent. The summation or integration being extended to the entire range of x.

$$\text{Thus, } M_0(t) = E(e^{tx}) = E\left[1 + tx + \frac{(tx)^2}{2!} + \text{....} + \frac{(tx)^r}{r!} + \text{.......}\right]$$

$$= 1 + t\,E(x) + \frac{t^2}{2!} E(x^2) + \text{.... .....} + \frac{t^r}{r!} E(x^r) \text{...}$$

$$= 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \text{.... .....} + \frac{t^r}{r!}\mu'_r + \text{....}$$

where $\mu'_r = E(x^r) = \sum_x x^r p(x)$; for discrete distribution,

$$= \int_{-\infty}^{\infty} x^r f(x)\, dx \text{ ; for continuous distribution.}$$

Thus the Co efficient of $\frac{t^r}{r!}$ in $M_o(t)$ gives μ'_r.

Since $M_o(t)$ generates moments, it is known as moment generating function (m. g. f.).

It is easy to varify that, $\mu'_r = \left.\frac{d^r M_o(t)}{dt^r}\right]_{t=0}$.

The moment generating function about the arithmetic mean μ is defined by

$M_\mu(t) = E[e^{t(x-\mu)}] = e^{-\mu t} E(e^{tx})$

$= e^{-\mu t} M_o(t).$ ......(7.11)

It can be easily varified as earlier that

$$\mu_r = \left.\frac{d^r M_\mu(t)}{dt^r}\right]_{t=0}.$$

A Property of Moment Generating Function : The moment generating function of the sum of a number of independent random variables is equal to the product of their respective moment generating functions.

Proof : Let $x_1, x_2, \dots x_n$ be n independent random variables (not the values of the variable x), then the moment generating functions of their sum $(x_1 + x_2 + \dots + x_n)$ with respect to origin is

$$M_o(t) = E\left[e^{t(x_1 + x_2 + \dots + x_n)}\right] = E[e^{tx_1} e^{tx_2} \dots\dots e^{tx_n}]$$

$$= E(e^{tx_1})E(e^{tx_2}) \dots\dots E(e^{tx_n}]$$

$$= M_o(t)_{x_1} M_o(t)_{x_2} \dots\dots M_o(t)_{x_n}$$

where $M_o(t)\, x_i$ indicates the m. g. f. of random variable x_i. Hence the theorem is proved.

7.6 Cumulant

The cumulant generating function k(t) is defined as

$k(t) = \log M_o(t).$ ...(7.12)

provided that right hand side can be expanded as a convergent series in power of t. If we expand k(t) in the following form

$$k(t) = k_1 t + k_2 \frac{t^2}{2!} + \dots\dots + k_r \frac{t^r}{r!} \dots\dots$$

then k_r = co-efficient of $\frac{t^r}{r!}$ is called the rth cumulant.

It is easy to varify that $k_r = \left.\frac{d^r k(t)}{dt^r}\right)_{t=0}$.

Relation Between Moments and Cumulants

We have $k(t) = k_1 t + k_2 \frac{t^2}{2!} + k_3 \frac{t^3}{3!} + \cdots\cdots$ ....(7.13)

Again $k(t) = \log M_o(t) = \log\left(1 + \mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \cdots\cdots\right)$

$$= \left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \cdots\cdots\right) - \frac{1}{2}\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \cdots\cdots\right)^2$$

$$+ \frac{1}{3}\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \cdots\cdots\right)^3 - \cdots\cdots \qquad \text{..........(7.14)}$$

Now equating the identical power of t of (7.13) and (7.14) we have

$k_1 = \mu_1'$

$k_2 = \mu_2' - \mu_1'^2 = \mu_2$

$k_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \mu_3$

$k_4 = \mu_4' - 3\mu_2'^2 - 4\mu_3'\mu_1' + 12\mu_2'\mu_1'^2 - 6\mu_1'^4$

$= \mu_4 - 3\mu_2^2$.

7.7 Characteristic Function

The characteristic function of a random variable x about origin is defined as

$\varphi_o(t) = E(e^{itx}) = \sum_x e^{itx} P(x)$; for discrete probability distribution.

$= \int_{-\infty}^{\infty} e^{itx} f(x)dx$; for continuous probability distribution. ....(7.15)

It can be easily shown that the rth moment about origin is given by

$$\mu_r' = \left.\frac{d^r \varphi_0(t)}{i^r\, dt^r}\right]_{t=0}$$

The characteristic function about the mean μ is given by

$$\varphi_\mu(t) = E[e^{it(x-\mu)}] = e^{-i\mu t}\,\varphi(t).$$

The rth central moment, μ_r is given by

$$\left[\frac{d^r \varphi_\mu(t)}{i^r dt^r}\right]_{t=0}$$

Example 7.7 Find the characteristic function of $f(x) = e^{-x}, 0 \le x \le \infty$ and hence find mean and variacne of f(x).

Solution : We know, $f(x) = e^{-x}, 0 \le x \le \infty$

$$\varphi_o(t) = \int_0^\infty e^{itx} f(x)\,dx = \int_0^\infty e^{itx} e^{-x} dx = \int_0^\infty e^{-x(1-it)}\,dx$$

Putting $x(1 - it) = z$ or, $x = \dfrac{z}{(1-it)}$ $\therefore dx = \dfrac{dz}{(1-it)}$

$$\therefore \varphi_o(t) = \frac{1}{(1-it)}\int_0^\infty e^{-z} dz = \frac{1}{(1-it)}.\ \text{Since} \int_0^\infty e^{-z} dz = \lceil 1 = 1.$$

Therefore the characteristic function of $f(x) = e^{-x}$ is $\varphi_o(t) = (1 - it)^{-1}$.

Now, $\dfrac{d\varphi_o(t)}{dt} = -(1-it)^{-2}(-i)$

$= i(1 - it)^{-2}$

$$\therefore \text{Mean} = \mu_1' = \left.\frac{d\varphi_o(t)}{idt}\right]_{t=0} = 1.$$

Again $\dfrac{d^2\varphi_o(t)}{dt^2} = -2i(1-it)^{-3}(-i)$

$= -2(1 - it)^{-3}$ Since $i^2 = -1$

$$\therefore \mu_2' = \left.\frac{d^2\varphi_o(t)}{i^2t^2}\right]_{t=0} = 2.$$

Therefore, variance $= \mu_2 = \mu_2' - \mu_1'^2 = 2 - 1 = 1$.

A property of Characteristic Function : The Characteristic function of the sum of n independent random variables is equal to the product of their respective characteristic functions, i. e.

$$\varphi_o(t)_{x_1 + x_2 + \ldots\ldots + x_n} = \varphi_o(t)_{x_1}.\ \varphi_o(t)_{x_2} \ldots\ldots \varphi_o(t)_{x_n}. \qquad \ldots.(7.16)$$

Proof : Let $x_1, x_2, \ldots\ldots x_n$ be n independent random variable (not the values of the variable x) then the characteristic function of their sum $(x_1+x_2+\ldots\ldots+x_n)$ with respect to origin is

$\varphi_o(t)x_1 + x_2 + \ldots\ldots + x_n = E[e^{it}(x_1 + x_2 + \ldots\ldots + x_n)]$

$= E(e^{itx_1})\ E(e^{itx_2}) \ldots \ldots E(e^{itx_n})$.

$= \varphi_o(t)x_1 \varphi_o(t)x_2 \ldots\ldots \varphi_o(t)x_n.$ Hence proved.

Remark : The converse of (7.16) is not necessarily true.

Advantages of Characteristic Function Over Moment Generating Function :

1) The characteristic function always exists but moment generating function may or may not exist.
2) The characteristic function determines the distribution function uniquely i. e. a necessary and sufficient condition for two distribution with p. d. f's f(x) and f(y) are identical if their characteristic functions $\varphi(t)_x$ and $\varphi(t)_y$ are identical.
3) Characteristic function follows the following necessary conditions.
   i) $\varphi(t)$ is continuous in t.
   ii) $\varphi(t)$ is defined for every value of t.
   iii) $\varphi(o) = 1$.
   iv) $\varphi(t)$ and $\varphi(-t)$ are conjugate quantities.
   v) $|\varphi(t)| \leq 1 \leq \varphi(o)$.

Theorem 7.5 Inversion Theorem (without proof) : If $\varphi(t)$ be the characteristic function and f(x) be the p. d. f. of a random variable x then

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\,\varphi(t)dt \qquad \ldots\ldots(7.17)$$

Example 7.8 Find that p. d. f. of the random variable x ; $-\infty \le x \le \infty$ for which $\varphi(t) = e^{-\frac{t^2\sigma^2}{2}}$.

Solution : Let f(x) be the p. d. f. of the random variable then,

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\,\varphi(t)dt$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\, e^{-\frac{t^2\sigma^2}{2}}\, dt$$

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

Let us put $t\sigma + \frac{ix}{\sigma} = y \quad \therefore dt = \frac{dy}{\sigma}$.

The range of y becomes $-\infty$ and ∞.

$$\therefore\ f(x) = \frac{1}{2\pi} e^{-\frac{x^2}{2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} \frac{dy}{\sigma}$$

$$= \frac{1}{2\pi\sigma} e^{-\frac{x^2}{2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy$$

$$= \frac{1}{2\pi\sigma} e^{-\frac{x^2}{2\sigma^2}} \sqrt{2\pi} = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{x^2}{2\sigma^2}} ;\ -\infty \le x \le \infty.$$

which is the p. d. f. of the random variable x.

7.8 Law of Large Number

Usually the estimates are made of an unknown quantity (parameter) by taking the average of a number of repeated measurements of the quantity, each of which may contain some error. Therefore, it is of certain interest to study the properties of the estimates. An initial enquiry is made concerning its behaviour as the number of measurement increases i. e. $\to\infty$. The problem that the estimates converge in some sense to the true value of the parameter can be formulated in the following ways :

Let x_n, n=1, 2 .....be a sequence of observations and $\overline{x}_n$ is the average of n observations then what are the conditions under which we can say that

$$\overline{x}_n \to \mu \text{ (parameter)} \qquad \text{......(7.18)}$$

in any one of the following modes of convergence.?

a) Weakly or in probability (written as $x_n \to c$) if, for every given $\varepsilon > 0$

$$\lim_{n\to\infty} P\{|x_n - c| > \varepsilon\} = 0 \qquad \text{.............(7.19)}$$

b) Strongly or almost surely (written as $\lim_{n\to\infty} x_n = c$ with probability 1 or $x_n \xrightarrow{a.s.} c$

$$\text{if } P\left\{\lim_{n\to\infty} x_n = c\right\} = 1. \qquad \text{...(7.20)}$$

c) In quadratic mean(written as $x_n \xrightarrow{q.m.} c$) if,

$$\lim_{n\to\infty} E(x_n - c)^2 = 0 \qquad \text{..........(7.21)}$$

We shall generalise the problem further and ask for the condition under which

$$\overline{x}_n - \mu_n \to 0 \qquad \text{...........(7.22)}$$

where μ_n, n=1, 2.....is a sequence of constant sought to be measured by the sequence of observations x_n, n=1, 2..... The law of large number holds if the convergence such as (7.18) or (7.22) takes place. When the convergence is "in probability" given in (7.19) we shall see that the weak law of large number (W. L. L. N) holds and when it is "with probability 1" or "almost surely" given in (7.20), the strong law of large number (S. L. L. N) holds.

Some of the important theorems of law of large numbers are given below :

1. Chebyshev's Theorem (W. L. L. N.) : Let $E(x_i) = \mu$, $V(x_i)=\sigma^2$ and $cov(x_i x_j) < 0$, $i < j$. Then .

$$\lim_{n\to\infty} \frac{\sigma^2}{n} = 0 \text{ implies that } \overline{x}_n \xrightarrow{p} \mu$$

where $\overline{x}_n$ is the mean of a series of n observations.

Proof : The proof of the above theorem can be done with the help of Chebyshev's inequality.

We consider x as a continuous random variable. Then by definition,

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

$$= \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)dx.$$

For the first integral, $x \le \mu - k\sigma \Rightarrow (\mu - x) \ge k\sigma$.

and for the third integral $x \ge \mu + k\sigma \Rightarrow (x-\mu) \ge k\sigma$.

Now dropping the middle term and replacing $(x-\mu)^2$ by the value obtained here, we get,

$$\sigma^2 \ge k^2\sigma^2 \int_{-\infty}^{\mu-k\sigma} f(x)dx + k^2\sigma^2 \int_{\mu+k\sigma}^{\infty} f(x)\,dx. \ge k^2\sigma^2 P\{\,|x-\mu| \ge k\sigma\}$$

$$\therefore P\{\,|x-\mu| \ge k\sigma\} \le \frac{1}{k^2}$$

With the help of this result we have in our case,

$$P\{\,|\overline{x_n} - \mu| \ge k'\} \le \frac{\sigma^2}{k'^2 n} \to 0 \left[\text{Since, } V(\overline{x}_n) = \frac{\sigma^2}{n}\right]$$

which implies that $\overline{x_n} \to \mu$.

2. Khinchin's Theorem (W. L. L. N) : Let x_n, n = 1, 2.....be independent and identically distributed (i.i.d.) and E (x_n) exists. Then,

$E(x_n) = \mu \le \infty$ implies that $\overline{x} \xrightarrow{p} \mu$.

3. Kolmogorov Theorem (S. L. L. N) : Let x_1, x_2..... be a sequence of i. i. d. variables. Then a necessary and sufficient condition that

$\overline{x_n} \xrightarrow{a\ s} \mu$ is that $E(x_i)$ exists and is equal to μ.

The proof of theorem No. 2 and 3 are beyond the scope of this text.

8. PROBABILITY DISTRIBUTIONS

8.1 Introduction

In this chapter we have discussed some of the important discrete and continuous probability distributions which are of special importance in theory and practice of statistics.

The names of the probability distributions discussed in this text, are as follows :

a) Discrete Distributions

1) Binomial. 2) Piosson. 3) Negative Binomial. 4) Geometric.
5) Hypergeometric. 6) Multinomial. 7) Uniform or Rectangular.

b) Continuous Distributions.

1) Uniform or Rectangular, 2) Normal. 3) Gamma. 4) Beta.
5) Exponential. 6) Cauchy. 7) Laplace.

8.2 Binomial Distribution

Let an experiment be repeated for n independent trials each with one of two possible outcomes, 'success' or 'failure'. The number of success, x in n trials is a discrete random variable which can assume values 0, 1, 2,........n. Let p be the probability of success and q be the probability of failure in a single trial so that $p + q = 1$. If the probability of success, p remains same from trial to trial, then the distribution of x is known as binomial distribution and its probability function is given by

$$p(x) = \binom{n}{x} p^x q^{n-x}; \quad x = 0, 1, 2, \ldots \ldots n \qquad \ldots.(8.1)$$

The binomial distribution was discovered by James Bernoulli (1654-1705) in the year 1700.

The following conditions must be statisfied for the binomial distribution.

i) There should be a fixed number of trials.
ii) The trials are independent.
iii) There are only two outcomes for each trial.
iv) The probability of success and hence the probability of failure remains same or constant from trial to trial.

Derivation : Let the first x trials resulted in success (S) and the rest (n-x) trials resulted in failure (F). Then the sequence of successes and failures be

$$\underbrace{S\ S.....S}_{x \text{ times}} \qquad \underbrace{F\ F.....F}_{(n-x) \text{ times}}$$

Since the trials are independent, the probability of this particular sequence is $p^x q^{n-x}$. But we are interest in any x trials being successes and since x trials can be chosen out of n in $\binom{n}{x}$ mutually exclusive ways, the probability p(x) of x successes is given by $p(x) = \binom{n}{x} p^x q^{n-x}$; $x = 0, 1, 2, ...n.$

The probability distribution function of the number of success, so attained is called the binomial probability distribution for the obvious reasons that the probabilities of 0, 1, 2,......n successes viz.

$q^n, \binom{n}{1} q^{n-1} p, \binom{n}{2} q^{n-2} p^2, ..., p^n$ are the successive terms of the binomial expansion $(q + p)^n$.

Remarks :

1) The probability function denoted by (8.1) satisfies the two properties of density function i. e.

   a) $p(x) = \binom{n}{x} p^x q^{n-x} \geq 0$ for all values of x,

   b) $\sum\limits_{x=0} p(x) = \sum\limits_{x=0} \binom{n}{x} p^x q^{n-x} = (q + p)^n = 1$; Since $p + q = 1$.

2). The two independent constants n and p of the distribution are known as the parameters of the distribution.

3) If $p = q = \frac{1}{2}$, the binomial distribution is symmetric otherwise it is skew.

Example 8.1 Four unbiased coins are tossed simultaniously. What is the probability of getting

a) exactly two heads? b) at least three heads?

Solution : The probability of getting x heads in a throw of 4 unbiased coins is

$$p(x) = \binom{4}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{4-x}; \quad x=0, 1, 2, 3, 4.$$

a) Probability of getting exactly two heads is given by

$$p(2) = \binom{4}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{2} = \frac{6}{16} = \frac{3}{8}$$

$16 \times \frac{1}{4} \times \frac{1}{4}$

b) Probability of getting at least three heads is given by

$$\text{Prob}\{x \geq 3\} = p(3) + p(4) = \binom{4}{3}\left(\frac{1}{2}\right)^{3}\frac{1}{2} + \left(\frac{1}{2}\right)^{4} = \frac{4}{16} + \frac{1}{16} = \frac{5}{16}.$$

Properties of Binomial Distribution :

Mean (μ) : We know, $\mu = \mu_1' = E(x) = \sum_{x=0}^{n} xp(x)$

$$= \sum_{x=0}^{n} x\binom{n}{x} p^x q^{n-x}$$

$$= np\sum_{x=1}^{n}\binom{n-1}{x-1} p^{x-1}q^{n-x}$$

$$= np\,(q+p)^{n-1} = np. \text{ Since } p+q=1. \qquad \text{.......(8.2)}$$

Also $\binom{n}{x} = \frac{n}{x}\binom{n-1}{x-1} = \frac{n(n-1)}{x(x-1)}\binom{n-2}{x-2}\ \frac{n(n-1)(n-2)}{x(x-1)(x-2)}\binom{n-3}{x-3}$

and so on.

Variance (σ^2): We know $\sigma^2 = \mu_2' - \mu_1'^2 = \mu_2$

where $\mu_2' = E(x^2) = E[x(x-1) + x]$

$$= E[x(x-1)] + E(x) \qquad \text{.............(8.3)}$$

Again $E[x(x-1)] = \sum_{x=0}^{n} x(x-1)\binom{n}{x} p^x q^{n-x}$

$$= \sum_{x=0}^{n} x(x-1)\frac{n(n-1)}{x(x-1)}\binom{n-2}{x-2} p^x q^{n-x}$$

$$=n(n-1)p^2\sum_{x=2}^{n}\binom{n-2}{x-2}p^{x-2}q^{n-x}$$

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

$$=n(n-1)p^2 \qquad \text{..........(8.4)}$$

We have already known $E(x) = np$ in (8.2).

Therefore, from (8.3) we get $\mu_2' = (n-1)p^2+np$.

Now, $\mu^2 = \mu_2' - \mu_1'^2$

$= n(n-1)p^2 + np - (np)^2$

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

$= np(1-p) = npq$, Since $1-p=q$. ....(8.5)

Third Moment (μ_3) : We know, $\mu'_3=E(x^3)=E[x(x-1)(x-2)+3x(x-1)+x]$

$=E[x(x-1)(x-2)]+3E[x(x-1)]+E(x)$ .....(8.6)

Now, $E[x(x-1)(x-2)] = \sum_{x=0}^{n} x(x-1)(x-2)\binom{n}{x}p^xq^{n-x}$

$$= \sum_{x=0}^{n} x(x-1)(x-2).\frac{n(n-1)(n-2)}{x(x-1)(x-2)}p^3\binom{n-3}{x-3}p^{x-3}q^{n-x}$$

$$= n(n-1)(n-2)p^3\sum_{x=3}^{n}\binom{n-3}{x-3}p^{x-3}q^{n-x}$$

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

$$= n(n-1)(n-2)p^3 \qquad \text{......(8.7)}$$

Form (8.2). (8.4) and (8.7), we get

$\mu'_3 = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$

Therefore, the third moment is

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

$=n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np - 3\{n(n-1)p^2 + np\}np + 2n^3p^3.$

$n^3p^3 - 3n^2p^3 + 2np^3 + 3n^2p^2 - 3np^2 + np - 3n^3p^3 + 3n^2p^3 - 3n^2p^2 + 2n^3p^3$

$=2np^3 - 3np^2 + np$

$=np[2p^2 - 3p + 1]$

$=np(1-p)(1-2p)$

$=npq(q-p)$ ....(8.8)

Fourth Moment (μ_4) : We know,

$\mu_4'=E(x^4) = E[x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x]$

$=E[x(x-1)(x-2)(x-3)] + 6E[x(x-1)(x-2)] + 7E[x(x-1)] + E(x)$ .....(8.9)

Now, $E[x(x-1)(x-2)(x-3)] = \sum_{x=0}^{n} x(x-1)(x-2)(x-3)\binom{n}{x}p^x q^{n-x}$

$= \sum_{x=0}^{n} x(x-1)(x-2)(x-3)\frac{n(n-1)(n-2)(n-3)}{x(x-1)(x-2)(x-3)}\binom{n-4}{x-4}p^x q^{n-x}$

$=n(n-1)(n-2)(n-3)p^4 \sum_{x=4}^{n}\binom{n-4}{x-4}p^{x-4}q^{n-x}$

$=n(n-1)(n-2)(n-3)p^4(p+q)^{n-4}$

$=n(n-1)(n-1)(n-3)p^4$ ....(8.10)

From (8.2), (8.4), (8.7) and (8.10), we have,

$\mu_4' = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np.$

Therefore the fourth moment (μ_4) is,

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

$=3n^2p^2q^2 + npq(1-6pq)$ [on simplification] ....(8.11)

$$\left.\begin{aligned}\text{Hence } \beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq}\\ \text{and} \beta_2 &= \frac{\mu_4}{\mu_2^2} = 3 + \frac{1-6pq}{npq}\end{aligned}\right\} \quad \text{........(8.12)}$$

Remarks : 1) The mean is alway greater than the variance as $q < 1$.

2) As the number of trials n increases infinitely,

$\beta_1 \to 0$ and $\beta_2 \to 3$.

Moment Generating Function of Binomial Distribution : The m. g. f. about origin of the binomial variate x is

$M(t) = E(e^{tx}) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x q^{n-x}$

$= \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x q^{n-x}$

$=(q+pe^t)^n$ ....(8.13)

Differentiating (8.13) with respect to t, we get,

$$\frac{dM(t)}{dt} = n(q+pe^t)^{n-1}pe^t \qquad \text{........(8.14)}$$

$$\therefore \mu_1' = \left.\frac{dM(t)}{dt}\right]_{t=0} = n(q+p)^{n-1}\ p = np$$

Again differentiating (8.14) with respect to t, we get,

$$\frac{d^2M(t)}{dt^2} = n(n-1)(q+pe^t)^{n-2}(pe^t)^2 + n(q+pe^t)^{n-1}pe^t.$$

$$= n(n-1)p^2(q+pe^t)^{n-2}e^{2t} + np(q+pe^t)^{n-1}e^t. \qquad \text{...(8.15)}$$

$$\therefore \mu_2' = \left.\frac{d^2M(t)}{dt^2}\right]_{t=0} = n(n-1)p^2 + np.$$

$$\therefore \mu_2 = \mu_2' - \mu_1'^2 = n(n-1)p^2 + np - n^2p^2 = npq.$$

Again differentiating (8.15) with respect to t, we get,

$$\frac{d^3M(t)}{dt^3} = n(n-1)(n-2)p^2(q+pe^t)^{n-3}pe^te^{2t} + n(n-1)p^2$$

$$(q+pe^t)^{n-2}2e^{2t} + (n-1)p(q+pe^t)^{n-2}pe^te^t + np(q+pe^t)^{n-1}e^t.$$

$$= n(n-1)(n-2)p^3(q+pe^t)^{n-3}e^{3t} + 3n(n-1)p^2(q+pe^t)^{n-2}e^{2t}$$

$$+ np(q+pe^t)^{n-1}e^t. \qquad \text{........(8.16)}$$

$$\therefore \mu_3 = \left.\frac{d^3M(t)}{dt^3}\right]_{t=0} = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$$

It can be easily shown that, $\mu_3 = npq(q-p)$. (after simplification).

Again differentiating (8.16) with respect to t we get,

$$\frac{d^4M(t)}{dt^4} = n(n-1)(n-2)(n-3)p^4(q+pe^t)^{n-4}e^{4t}$$

$$+ 6n(n-1)(n-2)p^3(q+pe^t)^{n-3}e^{3t} + n(n-1)p^2(q+pe^t)^{n-2}e^{2t}$$

$$+ np(q+pe^t)^{n-1}e^t.$$

$$\therefore \mu_4' = \left.\frac{d^4M(t)}{dt^4}\right]_{t=0} = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np$$

Therefore, it can be easily shown that the fourth moment,

$\mu_4 = 3n^2p^2q^2 + npq(1-6pq)$ (after simplification).

Characteristic Function of Binomial Distribution : The characteristic function about origin of a binomial variate x is

$$\varphi(t) = E(e^{itx}) = \sum_{x=0}^{n} e^{itx}\binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (pe^{it})^x q^{n-x} = (q + pe^{it})^n \qquad \text{..........(8.17)}$$

Differentiating $\varphi(t)$, once, twice etc. with respect to it and putting $t = 0$, we get the same results of μ_2, μ_3 and μ_4.

Recurrence Relation for the Probabilities of Binomial Distribution :

We know, $P(x) = \binom{n}{x} p^x q^{n-x}$ and $P(x+1) = \binom{n}{x+1} p^{x+1} q^{n-x-1}$

Now, $$\frac{P(x+1)}{P(x)} = \frac{\binom{n}{x+1} p^{x+1} q^{n-x-1}}{\binom{n}{x} p^x q^{n-x}} = \frac{n-x}{x+1}\,\frac{p}{q}$$

Hence, $P(x+1) = \frac{n-x}{x+1}\,\frac{p}{q}\,p(x)$, $x = 0, 1, 2 \ldots\ldots\ldots n.$

which is the required recurrence relation. This relation is helpful for calculating probabilities for different values of the binomial variate. The only probability, we need to calculate is p(o) which is equal to q^n. If p is not known, it can be estimated by $\hat{p} = \frac{\bar{x}}{n}$, where $\bar{x}$ is the sample mean of the distribution.

Example 8.2 Seven coins are tossed at a time and the number of head are noted. The experiment is repeated 128 times and the distribution is obtained on the next page.

No. of heads :	0	1	2	3	4	5	6	7
Frequencies :	7	6	19	35	30	23	7	1

Fit a binomial distribution to the above data assuming that,

i) the coin is unbiased i. e. $p = q = \frac{1}{2}$,

ii) the nature of the coin in not known i. e. p is unknown.

Solution : (i) Since $p = q = \frac{1}{2}$, $\frac{p}{q} = 1$ and $P(o) = \left(\frac{1}{2}\right)^7 = \frac{1}{128}$

From the recurrence relation P(1), P(2).........can be obtained as follows :

Table-8.1

x	$\frac{n-x}{x+1}\frac{p}{q}$	P(x)	E = N x P(x)
0	7	$\frac{1}{128}$	1
1	3	$\frac{7}{128}$	7
2	$\frac{5}{3}$	$\frac{21}{128}$	21
3	1	$\frac{35}{128}$	35
4	$\frac{3}{5}$	$\frac{35}{128}$	35
5	$\frac{1}{3}$	$\frac{21}{128}$	21
6	$\frac{1}{7}$	$\frac{7}{128}$	7
7	—	$\frac{1}{128}$	1
Total		1	128

(ii) Since p is not known, it can be estimated as follows :

We know, $\bar{x} = np = \frac{1}{N}\Sigma f_i x_i = \frac{433}{128} = 3.3828$ (app)

$\therefore$ p = 0.48326 and q = 0.51674

And $\frac{p}{q} = 0.93521$. $P(o) = (0.51674)^7 = .00984$.

Table-8.2

x	$\frac{n-x}{x+1}\frac{p}{q}$	P(x)	* E = N x P(x).
0	6.54647	0.00984	$1.25952 \approx 1$
1	2.80563	0.06440	$8.2432 \approx 8$
2	1.55868	0.18069	$23.12832 \approx 23$
3	0.93521	0.28164	$36.04992 \approx 36$
4	0.56113	0.26339	$33.71392 \approx 34$
5	0.31174	0.14779	$18.9171 \approx 19$
6	0.13360	0.04607	$5.896 \approx 6$
7	—	0.00618	$0.791 \approx 1$
		1	128

* Since the number of trials cannot be fraction, we converted the expected values into nearest integers.

8.3 Poisson Distribution

The poisson distribution was discovered by S. Devis Poisson (1781-1840) in the year 1837.

Poisson distribution can be defined as the limiting case of the binomial distribution under the following conditions :

i) the number of trials are very large i. e. $n \to \infty$,

ii) the probability of success, p is very small i. e. $p \to 0$ and

iii) the mean of the binomial distribution np = m, a finite and positive constant.

The probability function of poisson distribution is given by

$$p(x) = \frac{e^{-m}m^x}{x!};\ x = 0, 1, \ldots\ldots\infty. \qquad \ldots\ldots(8.18)$$

Derivation of Poisson Distribution from Binomial Distribution : The probability of x success in a series of n independent trail given in (8.1) is given by, $p(x) = \binom{n}{x} p^x q^{n-x}$; 0, 1, 2, ...n.

We want the limiting form of p(x) under the above three conditions.

We have, $p(x) = \binom{n}{x} p^x(1-p)^{n-x}$

$$= \frac{n!}{x!(n-x)!}\left(\frac{m}{n}\right)^x\left(1-\frac{m}{n}\right)^{n-x} \quad \text{Since, } n\,p = m.$$

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

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

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

As $n\rightarrow\infty$; $\frac{1}{n}$, $\frac{2}{n}$ etc. tend to zero, $\left(1-\frac{m}{n}\right)^x$ tends to 1 and $\left(1-\frac{m}{n}\right)^n$ tends to e^{-m}

Therefore, $\underset{n\rightarrow\infty}{\mathrm{Lt}}\ p(x) = \frac{m^x}{x!}e^{-m}$ for fixed x and x=0, 1, 2, ...∞ which is the required probability function of the poisson distribution.

Remarks :

1) It should be noted that $\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty}\frac{e^{-m}m^x}{x!} = e^{-m}e^{m} = 1.$

2) m is the only parameter of the distribution and $m > 0$.
3) Following are some examples of poisson variates.
   i) Number of suicides reported in a particular city within 10 years (say).
   ii) Number of air accidents in some unit of time.
   iii) Number of telephone calls received at a particular telephone exchange in some unit of time.

Example 8.3 A manufacturer of pins knows that 5% of his product is defective. If he sells pins in boxes of 100 and guarentees that not more than 10 pins will be defective. What is the approximate probability that a box will fail to meet the guaranteed quality?

Solution : We have given n=10, Probability of getting defective pins, p = .05.

Therefore, m = mean number of defective pins ; np = 100 x .05 = 5.

Since p is very small, we may use poisson distribution. Probability of x defective pins in a box of 100 pins is

$$p(x) = \frac{e^{-m}m^x}{x!} = \frac{e^{-5}5^x}{x!};\ x = 0, 1, 2, \ldots\ldots$$

Probability that a box will fail to meet the guaranteed quality is

$$p(x > 10) = 1 - P(x \le 10) = 1 - \sum_{x=0}^{10} \frac{e^{-5}5^x}{x!}$$

$$= 1 - e^{-5} \sum_{x=0}^{10} \frac{5^x}{x!}$$

Properties of Poisson Distribution :

Mean (μ) : $\mu = \mu_1' = E(x) = \sum_{x=0}^{\infty} xp(x).$

$$= \sum_{x=0}^{\infty} x \frac{e^{-m}m^x}{x!} = me^{-m} \sum_{x=1}^{\infty} \frac{m^{x-1}}{(x-1)!}$$

$$= me^{-m}e^{m} = m. \qquad \ldots\ldots\ldots(8.19)$$

Hence the mean of poisson distribution is m.

Variance (σ^2) :

$\mu_2' = E(x^2) = E[x(x-1) + x]$

$$= E[x(x-1)] + E(x) \qquad \ldots\ldots\ldots\ldots(8.20)$$

$$\text{Now, } E[x(x-1)] = \sum_{x=0}^{\infty} x(x-1) \frac{e^{-m}m^x}{x!}$$

$$= m^2e^{-m} \sum_{x=2}^{\infty} \frac{m^{x-2}}{(x-2)!}$$

$$= m^2e^{-m}e^{m} = m^2 \qquad \ldots\ldots\ldots(8.21)$$

From (8.19), (8.20) and (8.21) we have

$\mu_2' = m^2 + m$

Therefore, the variance, $\sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = m^2 + m - m^2 = m.$

Third moment (μ_3) :

We know, $\mu_3' = E(x^3) = E[x(x-1)(x-2) + 3x(x-1) + x]$

$$= E[x(x-1)(x-2)] + 3E[x(x-1)] + E(x) \quad \text{....(8.22)}$$

$$\text{Now, } E[x(x-1)(x-2)] = \sum_{x=0}^{\infty} x(x-1)(x-2)\frac{e^{-m}m^x}{x!}$$

$$= m^3e^{-m}\sum_{x=3}^{\infty}\frac{m^{x-3}}{(x-3)!} = m^3e^{-m}e^m = m^3 \quad \text{...(8.23)}$$

From (8.19), (8.21), (8.22) and (8.23) we have

$\mu_3' = m^3 + 3m^2 + m$

Therefore the third moment is, $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$

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

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

Fourth moment (μ_4) :

We know, $\mu_4' = E(x^4) = E[x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x]$

$$= E[x(x-1)(x-2)(x-3)] + 6E[x(x-1)(x-2)] + 7E[x(x-1)] + E(x). \quad \text{....(8.24)}$$

$$\text{Now, } E[x(x-1)(x-2)(x-3)] = \sum_{x=0}^{\infty} x(x-1)(x-2)(x-3)\frac{e^{-m}m^x}{x!}$$

$$= m^4e^{-m}\sum_{x=4}^{\infty}\frac{m^{x-4}}{(x-4)!} = m^4e^{-m}e^m = m^4 \quad \text{........(8.25)}$$

From (8.19), (8.21), (8.23), (8.24) and (8.25) we have

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

Therefore the fourth moment is, $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4.$

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

$= 3m^2 + m.$

$$\left.\begin{aligned} \text{Hence } \beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{1}{m} \\ \text{and } \beta_2 &= \frac{\mu_4}{\mu_2^2} = 3 + \frac{1}{m} \end{aligned}\right\} \quad \text{........(8.26)}$$

Remarks :

1) Mean and variance of poisson distribution are each equal to m. This is an important characteristic of this distribution.

2) As $m \to \infty$; $\beta_1 \to 0$ and $\beta_2 \to 3$.

Moment Generating Function of Poisson Distribution :
The m. g. f. about origin of a poisson variate x is

$$M(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-m} m^x}{x!} = e^{-m} \sum_{x=0}^{\infty} \frac{(me^t)^x}{x!}$$

$$= e^{-m} e^{me^t} = e^{m(e^t - 1)}. \qquad \text{............(8.27)}$$

Now differentiating M(t) with respect to, t, we get

$$\frac{dM(t)}{dt} = e^{m(e^t - 1)} me^t = m\, e^{m(e^t - 1)} e^t \qquad \text{........(8.28)}$$

$$\therefore \mu_1' = \left.\frac{dM(t)}{d(t)}\right]_{t=0} = m.$$

Again differentiating (8.28) with respect to t, we get

$$\frac{d^2M(t)}{dt^2} = e^{m(e^t - 1)} (me^t)^2 + e^{m(e^t - 1)} me^t.$$

$$= m^2 e^{m(e^t - 1)} e^{2t} + me^{m(e^t - 1)} e^t \qquad \text{......(8.29)}$$

$$\therefore \mu_2 = \left.\frac{d^2M(t)}{dt^2}\right]_{t=0} = m^2 + m.$$

Therefore, $\sigma^2 = \mu_2 = \mu'_2 - \mu_1'^2 = m^2 + m - m^2 = m.$
Again differentiating (8.29) with respect to t we get,

$$\frac{d^3M(t)}{dt^3} = m^2 e^{m(e^t - 1)} (me^t) e^{2t} + m^2 e^{m(e^t - 1)} 2e^{2t} + me^{m(e^t - 1)} (me^t) e^t + me^{m(e^t - 1)} e^t$$

$$\text{......(8.30)}$$

$$= m^3 e^{m(e^t - 1)} e^t + 2m^2 e^{m(e^t - 1)} e^{2t} + m^2 e^{m(e^t - 1)} e^{2t} + me^{m(e^t - 1)} e^t.$$

$$\therefore \mu_3' = \left.\frac{d^3M(t)}{dt^3}\right]_{t=0} = m^3 + 2m^2 + m^2 + m = m^3 + 3m^2 + m.$$

It can be easily shown that $\mu_3 = m$.
Once again differentiating (8.30) with respect to t we get,

$$\frac{d^4M(t)}{dt^4} = m^3e^{m(e^t-1)}me^te^{3t} + m^3e^{m(e^t-1)}3e^t + 3m^2e^{m(e^t-1)}me^te^{2t}$$
$$+ 3m^2e^{m(e^t-1)}2e^{2t} + me^{m(e^t-1)}me^te^t + me^{m(e^t-1)}e^t.$$

$$\therefore \mu'_4 = \left.\frac{d^4M(t)}{dt^4}\right]_{t=0} = m^4 + 6m^3 + 7m^2 + m$$

Therefore, $\mu_4 = 3m^2 + m$. On simplification.

Characteristic Function of Poisson Distribution :

The characteristic function of a poisson variate x is,

$$\varphi(t) = E(e^{itx}) = \sum_{x=0}^{\infty} e^{itx}p(x)$$

$$= \sum_{x=0}^{\infty} e^{itx}\frac{e^{-m}m^x}{x!} = e^{-m}\sum_{x=0}^{\infty}\frac{(me^{it})^x}{x!}$$

$$= e^{-m}e^{me^{it}} = e^{m(e^{it}-1)} \qquad \text{........(8.30)}$$

Differentiating $\varphi(t)$ once, twice etc with respect to it and putting t=0 we get the same value of μ_2, μ_3 and μ_4.

Additive Property of Independent Poisson Variates :

If two independent poisson varaites x_1 and x_2 have mean m_1 and m_2 respectively, then their sum $y=x_1 + x_2$ is also a poisson variate with mean m_1+m_2.

Proof : Let $M_1(t)$ and $M_2(t)$ be the moment generating functions of poisson variates x_1 and x_2 respectively and M(t) be the moment generating funtion of their sum, then

$M_1(t) = e^{m_1(e^t-1)}$ and $M_2(t) = e^{m_2(e^t-1)}$

Since x_1 and x_2 are independent,

$$M(t) = E[e^{t(x_1+x_2)}] = E[e^{tx_1}e^{tx_2}]$$

$$= E(e^{tx_1})\,E(e^{tx_2}) = M_1(t)\,M_2(t)$$

$$= e^{m_1(e^t-1)}e^{m_2(e^t-1)} = e^{(m_1+m_2)(e^t-1)}$$

which is the moment generating function of y indicating a poisson variate with mean (m_1+m_2). Hence proved.

Recurrence Relation for the Probabilities of Poisson Distribution :

We know, $p(x) = \frac{e^{-m}m^x}{x!}$ and $p(x+1) = \frac{e^{-m}m^{x+1}}{(x+1)!}$

Now, $\frac{p(x+1)}{p(x)} = \frac{e^{-m}m^{(x+1)}}{(x+1)!} \times \frac{x!}{e^{-m}m^x} = \frac{m}{x+1}$

Hence, $p(x+1) = \frac{m}{x+1} p(x), \quad x = 0, 1, 2 \ldots\ldots$

which is the required recurrence relation. This relation is helpful for calculating probabilities for different values of poisson variate. The only probability, we need to calculate is p(o), which is equal to e^{-m}, where m is the mean of the distribution, if m is not known it can be estimated from the given data.

Example 8.4 The following data show the suicides of 1096 women in 8 cities in a country during 14 years.

No. of suicides	0	1	2	3	4	5	6	7
Frequency	364	376	218	89	33	13	2	1

Fit a poisson distribution to the above data.

Solution : Since m is not known, it can be estimated as follows :

$\hat{m} = \bar{x} = \frac{1}{N}\Sigma f_i x_i = \frac{1295}{1096} = 1.18$ Therefore, $p(o) = e^{-m} = e^{-1.18} = .30728$ (app).

Table-8.3

x	$\frac{m}{x+1}$	P(x)	*E = N x P(x).
0	1.1800	0.30728	$336.8 \approx 337$
1	0.5 900	0.36259	$397.5 \approx 398$
2	0.3933	0.21393	$234.5 \approx 235$
3	0.2950	0.08414	$92.2 \approx 92$
4	0.2360	0.02482	$27.2 \approx 27$
5	0.1967	0.00585	$6.4 \approx 6$
6	0.1686	0.00115	$1.3 \approx 1$
7	—	0.00023	$0.3 \approx 0$
		1	1096

* Since the number of suicides cannot be fraction we converted the expected values into nearest integers.

8.4 Negative Binomial Distribution

The equality of the mean and variance is an important characteristic of the poisson distribution whereas for the binomial distribution the mean is always greater than the variance. But its opposite feature that the variance is greater than the mean is seen in negative binomial distribution. The negative binomial distribution has been found to occur in many biological situations and can come about as a result of clustering (or contagian) among the successes of an otherwise binomial population e. g. death of insects, number of insect bites per apple etc.

A random variable x is said to follow a negative binomial distribution if its probability function is given by.

$$p(x) = \binom{x+r-1}{r-1} p^r q^x; \quad x = 0, 1, 2, \ldots\text{and } r > 0, \qquad \ldots\ldots\ldots(8.31)$$

where p is the probability of success and $p+q = 1$.

Derivation of Negative Binomial Distribution :

Let p (x) be the probability that there are x failure, preceeding the rth success in (x+r) trials. Here the trials are independent and the probability of success p in a trial remains constant from trial to trial. Clearly the last trial must be a success whose probability is p. In the remaining (x+r-1) trials, we must have (r-1) successes whose probability is given by

$$\binom{x+r-1}{r-1} p^{r-1}q^x \qquad \ldots\ldots\ldots(8.32)$$

Hence multiplying the two probabilities we get,

$$p(x) = \binom{x+r-1}{r-1} p^r q^x; \quad x = 0, 1, \ldots\ldots\text{and } r > 0$$

We know, $\binom{x+r-1}{r-1} = \binom{x+r-1}{x}$ $\left[\text{Since, } \binom{n}{r} = \binom{n}{n-r}\right]$

$$= \frac{(x+r-1)(x+r-2)\ldots(r+1)r}{x!}$$

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

$$= (-1)^x \binom{-r}{x}$$

Therefore (8.31) reduces to $p(x) = \binom{-r}{x} p^r(-q)^x ; \quad x = 0, 1, 2, \ldots\ldots\ldots\ldots\ldots$ (8.33)

which is the (x + 1)th term in the expension of $p^r(1 - q)^{-r}$, a binomial expansion with a negative index. Hence the distribution is known as negative binomial distribution.

Remarks :

1. The assignment of probability is permissible since,

$$\sum_{x=0}^{\infty} p(x) = p^r \sum_{x=0}^{\infty} \binom{-r}{x} (-q)^x = p^r(1-q)^{-r} = 1.$$

2. The distribution contains two parameters p and r.

Properties of Negative Binomial Distribution :

Mean (μ) : $\mu = \mu_1' = E(x) = \sum_{x=0}^{\infty} x \binom{-r}{x} p^r(-q)^x$

$$= p^r(-q) \sum_{x=0}^{\infty} x.\frac{-r}{x}.\binom{-r-1}{x-1} (-q)^{x-1}$$

$$= p^r(-q)(-r) \sum_{x=1}^{\infty} \binom{-r-1}{x-1} -q^{x-1}$$

$$= rqp^r(1-q)^{-r-1} = rqp^r p^{-r-1} = \frac{rq}{p} \qquad \ldots\ldots\ldots(8.34)$$

Variance (σ^2) :

We know $\mu_2' = E(x^2) = E[x(x-1) + x]$

$$= E[x(x-1)] + E(x) \qquad \ldots\ldots\ldots\ldots\ldots(8.35)$$

Now, $E[(x-1)x] \sum_{x=0}^{\infty} x(x-1)p(x)$

$$= \sum_{x=0}^{\infty} x(x-1) \binom{-r}{x} p^r(-q)^x$$

$$= p^r(-q)^2 \sum_{x=0}^{\infty} x(x-1)\frac{-r}{x}.\frac{-r-1}{x-1}\binom{-r-2}{x-2} (-q)^{x-2}$$

$$= p^r q^2 (-r)(-r-1) \sum_{x=2}^{\infty} \binom{-r-2}{x-2} (-q)^{x-2}$$

$$= r(r+1) p^r q^2 (1-q)^{-r-2}$$

$$= \frac{r(r+1)p^r q^2}{p^{r+2}} = \frac{r(r+1)q^2}{p^2} \qquad \text{..........(8.36)}$$

From (8.34), (8.35) and (8.36) we have,

$$\mu_2' = \frac{r(r+1)q^2}{p^2} + \frac{rq}{p}$$

$$\therefore \mu_2 = \mu_2' - \mu_1'^2 = \frac{r(r+1)q^2}{p^2} + \frac{rq}{p} - \frac{r^2q^2}{p^2} = \frac{rq}{p^2}.$$

Remark : In this case, mean is less than variance which is a distinguishing feature of this distribution.

Moment Generating Function of Negative Binomial Distribution :

The m. g. f. about origin of a negative binomial variate x is

$$M(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx} \binom{-r}{x} p^r (-q)^x$$

$$= p^r \sum_{x=0}^{\infty} \binom{-r}{x} (-qe^t)^x = p^r (1-qe^t)^{-r}. \qquad \text{..........(8.37)}$$

Differentiating M(t) with respect to t we get,

$$\frac{dM(t)}{dt} = p^r(-r)(1-qe^t)^{-r-1}(-qe^t)$$

$$= rqp^r e^t (1-qe^t)^{-r-1} \qquad \text{........(8.38)}$$

$$\mu_1' = \left.\frac{dM(t)}{dt}\right]_{t=0} = rqp^r(1-q)^{-r-1} = rqp^r p^{-r-1} = \frac{rq}{p}$$

Again differentiating (8.38) with respect to t we get,

$$\frac{d^2M(t)}{dt^2} = rqp^r(-r-1)(1-qe^t)^{-r-2}(-qe^t)e^t + rqp^r(1-qe^t)^{-r-1}e^t$$

$$= r(r+1)p^r q^2 (1-qe^t)^{-r-2} e^{2t} + rqp^r(1-qe^t)^{-r-1} e^t$$

$$\therefore \mu_2' = \left.\frac{d^2M(t)}{dt^2}\right]_{t=0} = r(r+1)q^2 p^r (1-q)^{-r-2} + rqp^r(1-q)^{-r-1}$$

$$=\frac{r(r+1)q^2}{p^2}+\frac{rq}{p}.$$

$\therefore \mu_2=\mu_2'-\mu_1'^2=\frac{rq}{p^2}$, on simplification.

Third and Fourth Moments : Following the method used in binomial and poisson distribution for the calculation of third and fourth moments we can easily show that the third moment, fourth moment, β_1 and β_2 of the negative binomial distribution are as follows :

$$\mu_3=\frac{rq(1+q)}{p^3} \text{ and } \mu_4=\frac{rq[p^2+3q(r+2)]}{p^4}$$

$$\therefore \beta_1=\frac{\mu_3^2}{\mu_2^3}=\frac{(1+q)^2}{rq} \text{ and } \beta_2=\frac{\mu_4}{\mu_2^2}=\frac{p^2+3q(r+2)}{rq}.$$

Poisson Distribution as a Limiting Case of Negative Binomial Distribution : Negative binomial distribution tends to poisson distribution as $r\to\infty$ and mean $=\frac{rq}{p}=m$ (a finite number).

We have, $m=\frac{rq}{p}$ or, $p=\frac{r}{m}q=\frac{r}{m}(1-p)$

or, $p\left(1+\frac{r}{m}\right)=\frac{r}{m}$ or, $p=\frac{r}{m+r}$. Hence $q=\frac{m}{m+r}$.

The probability function of a negative binomial variate x is

$$p(x)=\binom{x+r-1}{r-1}p^rq^x$$

$$\therefore \underset{r\to\infty}{Lt}\ p(x)=\underset{r\to\infty}{Lt}\binom{x+r-1}{r-1}\left(\frac{r}{m+r}\right)^r\left(\frac{m}{m+r}\right)^x$$

$$=\frac{(x+r-1)(x+r-2)\ldots\ldots(r+1)r}{x!}\underset{r\to\infty}{Lt}\left(\frac{1}{1+\frac{m}{r}}\right)^r\frac{m^x}{x^r\left(1+\frac{m}{r}\right)^x}$$

$$=\frac{m^x}{x!}\underset{r\to\infty}{Lt}.\left(1+\frac{x-1}{r}\right)\left(1+\frac{x-2}{r}\right)\ldots\ldots\left(1+\frac{1}{r}\right)\left(1+\frac{m}{r}\right)^{(r+x)}$$

$$=\frac{m^x}{x}\underset{r\to\infty}{Lt}e^{-m\frac{(r+x)}{r}}=\frac{m^x}{x!}e^{-m}$$

$$\therefore \underset{r \to \infty}{Lt}\; p(x) = \frac{e^{-m}m^x}{x!}$$

which is the p(x) of poisson variate with parameter m.

8.5 Geometric Distribution

A random variable x is said to have a geometric distribution if its probability function is given by

$$p(x) = pq^x \;;\; x = 0, 1, 2........ \qquad ....(8.39)$$

where p is the probability of success and $p + q=1$.

Derivation of Geometric Distribution : Let p(x) be the probability that there are x failures preceeding the first success in a series of independent trials. Let the probability of success in a trial is p which remains same from trial to trial. Then clearly,

$$p(x) = q^x p \;;\; x = 0, 1, 2.........$$

Remarks :

(1) Since the various probabilities for x = 0, 1, 2.........are the various terms of the geometric progression. Hence the name of the distribution is geometric distribution.

(2) Clearly, assignment of probability is permissible.

since $\sum\limits_{x=0}^{\infty} p(x) = \sum\limits_{x=0}^{\infty} q^x p = p(1-q)^{-1} = 1.$

(3). If we take r=1 in (8.31), the probability function of the negative binomial distribution, reduces to, $p(x) = q^x p \;;\; x = 0, 1, 2........$

which is the probability function of the geometric distribution. Hence negative binomial distribution may be regarded as the generalisation of the geometric distribution.

Properties of Geometric Distribution :

Mean (μ) : $\mu = \mu_1' = E(x) = \sum\limits_{x=0}^{\infty} xp(x) = \sum\limits_{x=0}^{\infty} xq^x p.$

$$= pq\sum_{x=1}^{\infty} xq^{x-1} = pq(1-q)^{-2} = \frac{q}{p} \qquad ....(8.40)$$

Variance (σ^2) :

We know, $\mu_2' = E(x^2) = E[x(x-1)] + E(x)$ ....(8.41)

Now, $E[x(x-1)] = \sum_{x=0}^{\infty} x(x-1)q^x p$

$$= 2pq^2 \sum_{x=2}^{\infty} \frac{x(x-1)}{2!} q^{x-2} = 2pq^2(1-q)^{-3} = \frac{2q^2}{p^2} \qquad \text{........(8.42)}$$

Therefore, from (8.40), (8.41) and (8.42) we have variance,

$$\sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = \frac{2q^2}{p^2} + \frac{q}{p} - \frac{q^2}{p^2} = \frac{q^2}{p^2} + \frac{q}{p} = \frac{q}{p^2}.$$

Moment Generating Function of Geometric Distribution :

The m. g. f. about origin of geometric distribution is,

$$M(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx}q^x p = p \sum_{x=0}^{\infty} (qe^t)^x = p(1-qe^t)^{-1} \qquad \text{..........(8.43)}$$

$$\therefore \mu_1' = \left.\frac{dM(t)}{dt}\right]_{t=0} = pq(1-qe^t)^{-2}\Big]_{t=0} = pq(1-q)^{-2} = \frac{q}{p}.$$

and $\mu_2 = \left.\frac{d^2M(t)}{dt^2}\right]_{t=0} = \frac{2q^2}{p^2} + \frac{q}{p}$ (on simplification).

$$\text{Therefore, } \mu_2 = \mu_2' - \mu_1'^2 = \frac{2q^2}{p^2} + \frac{q}{p} - \frac{q^2}{p^2} = \frac{q}{p^2}$$

Hence the mean and the variance of the geometric distribution are $\frac{q}{p}$ and $\frac{q}{p^2}$ respectively obtained by both the methods.

8.6 Hyper-geometric Distribution

The distribution is so termed as the moment generating function can be expressed in terms of hyper-geometric function.

When the population is finite and the sampling is done without replacement, we obtain hyper-geometric distribution.

Suppose r balls are drawn one at a time without replacement from a bag containing m white and n black balls. Then the probability of getting x white balls out of r is given by,

$$p(x) = \frac{\binom{m}{x}\binom{n}{r-x}}{\binom{m+n}{r}}; \quad x = 0, 1, 2 \ldots\ldots r \quad \begin{matrix} r \le m \\ r \le n \end{matrix} \qquad \ldots\ldots(8.44)$$

Remarks :

1) m, n, and r are known as the three parameters of **hyper-geometric** distribution.

2) The assignment of probability is permissible

Since $\sum_{x=0}^{r} \binom{m}{x}\binom{n}{r-x} / \binom{m+n}{r} = 1.$

Comparing the co - efficients of x^r in $(1 + x)^m(1 + x)^n = (1 + x)^{m+n}$

we get, $\sum_{x=0}^{r} \binom{m}{x}\binom{n}{r-x} = \binom{m+n}{r}.$

Properties of Hyper-geometric Distribution :

Mean $(\mu) : \mu = \mu_1' = E(x) = \sum_{x=0}^{r} xp(x)$

$$= \sum_{x=0}^{r} x \binom{m}{x}\binom{n}{r-x} / \binom{m+n}{r}.$$

$$= \sum_{x=0}^{r} x \frac{m}{x}\binom{m-1}{x-1}\binom{n}{r-x} / \binom{m+n}{r}.$$

$$= \frac{m}{\binom{m+n}{r}} \sum_{x=1}^{r} \binom{m-1}{x-1}\binom{n}{r-x}$$

$$= \frac{m}{\binom{m+n}{r}} \binom{m+n-1}{r-1} = \frac{mr}{m+n} \qquad \ldots\ldots\ldots(8.45)$$

Variance (σ^2) : We know, $\mu_2' = E(x^2) = E[x(x-1) + x]$

$$= E[x(x-1)] + E(x) \qquad \ldots\ldots\ldots(8.46)$$

Now, $E[x(x-1)] = \sum_{x=0}^{r} x(x-1)\,p(x)$

$$= \sum_{x=0}^{r} x(x-1)\frac{\binom{m}{x}\binom{n}{r-x}}{\binom{m+n}{r}}$$

$$= \sum_{x=0}^{r} x(x-1)\frac{\frac{m}{x}\cdot\frac{(m-1)}{(x-1)}\binom{m-2}{x-2}\binom{n}{r-x}}{\binom{m+n}{r}}$$

$$= \frac{m(m-1)}{\binom{m+n}{r}}\sum_{x=2}^{r}\binom{m-2}{x-2}\binom{n}{r-x}$$

$$= \frac{m(m-1)}{\binom{m+n}{r}}\binom{m+n-2}{r-2}$$

$$= \frac{m(m-1)\,r(r-1)}{(m+n)(m+n-1)} \qquad \text{.............(8.47)}$$

Therefore, from (8.45), (8.46) and (8.47) we have,

$$\mu_2' = \frac{mr(m-1)(r-1)}{(m+n)(m+n-1)} + \frac{mr}{(m+n)}$$

Hence the variance, $\sigma^2 = \mu_2 = \mu'_2 - \mu_1'^2$

$$= \frac{mr(m-1)(r-1)}{(m+n)(m+n-1)} + \frac{mr}{(m+n)} - \frac{m^2r^2}{(m+n)^2}$$

$$= \frac{mnr(m+n-r)}{(m+n)^2(m+n-1)} \text{ (on simplification).}$$

8.7 Multinomial Distribution

This distribution can be regarded as the generalisation of binomial distribution.

Let $E_1, E_2 ......E_r$ be r mutually exclusive and exhaustive outcomes of a trial with respective probabilities $p_1, p_2 ......p_r$, where $p_1 + p_2 + ...+ p_r = 1$.

The probability that n trials will result in E_1 occuring x_1 times, E_2 occuring x_2 times.......E_r occuring x_r times in a fixed definite order is

$$p_1^{x_1} \; p_2^{x_2} \cdots\cdots\cdots p_r^{x_r} \; ; \; \Sigma x_i = n.$$

But we are interested in events occuring in any order. The number of mutually exclusive ways in which this can happen is $\frac{n!}{x_1!\, x_2! \ldots\ldots x_r!}$.

Hence the required probability is

$$p(x_1, x_2 \ldots\ldots x_r) = \frac{n!}{x_1!\, x_2! \ldots\ldots\ldots x_r!} \; p_1^{x_1} \; p_2^{x_2} \cdots\cdots\cdots p_r^{x_r} \; ; \; 0 \le x_i \le n$$

This distribution is called multinomial probability distribution as the expression is the general term of the multinomial expansion of $(p_1+p_2+\ldots\ldots+p_r)^n$.

Moment Generating Function of Multinomial Distribution :

The moment generating function is given by

$$M(t) = M(t_1, t_2, \ldots\ldots\ldots t_r) = E(e^{t_1x_1 + t_2x_2 + \ldots\ldots + t_rx_r})$$

$$= \sum e^{t_1x_1 + t_2x_2 + \ldots\ldots + t_rx_r} \frac{n!}{x_1!x_2! \ldots\ldots x_r!} \; p_1^{x_1} \; p_2^{x_2} \cdots\cdots\cdots p_r^{x_r}$$

$$= \sum \frac{n!}{x_1!\, x_2! \ldots\ldots x_r!} (p_1e^{t_1})^{x_1} (p_2e^{t_2})^{x_2} \ldots\ldots (p_re^{t_r})^{x_r}$$

$$= (p_1e^{t_1} + p_2e^{t_2} + \ldots\ldots + p_re^{t_r})^n$$

Mean (μ_1) :

$$\mu_{1i} = \mu_1{}'_i = E(x_i) = \left.\frac{dM(t)}{dt_i}\right]_{t_1 = t_2 = \ldots\ldots = t_r = 0}$$

$$= np_ie^{t_i}(p_1e^{t_1} + p_2e^{t_2} + \ldots\ldots + p_re^{t_r})^{n-1}.$$

$$= np_i.$$

Variance (σ^2).

We have, $\mu_2{}' = E(x_i^2) = \left.\frac{d^2M(t)}{dt_i^2}\right]_{t_1 = t_2 = \ldots\ldots = t_r = 0}.$

We know,

$$\frac{d^2M(t)}{dt^2} = n(n-1)p_i^2e^{2t_i}\left(p_1e^{t_1} + p_2e^{t_2} + \dots\dots + p_re^{t_r}\right)^{n-2}$$

$$+ np_ie^{t_i}\left(p_1e^{t_1} + p_2e^{t_2} + \dots p_re^{t_r}\right)^{n-1}$$

$$\therefore \left.\frac{d^2M(t)}{dt_i^2}\right]_{t_1 = t_2 = \dots = t_r = 0} = n(n-1)p_i^2 + np_i$$

$\therefore$ Variance $(\sigma^2) = \mu_2 = \mu_2' - \mu_1'^2 = n(n-1)p_i^2 + np_i - n^2p_i^2 = np_i(1-p_i)$; $i=1, 2, \dots r$.

$$E(x_ix_j) = \left.\frac{d^2M(t)}{dt_idt_j}\right]_{\substack{i\neq j \\ t_1 = t_2 = \dots = t_r = 0.}}$$

$$\therefore \frac{d^2M(t)}{dt_idt_j} = np_ie^{t_i}(n-1)p_je^{t_j} + \left(p_1e^{t_1} + p_2e^{t_2} + \dots\dots + p_re^{t_r}\right)^{n-2}$$

$$\therefore \left.\frac{d^2M(t)}{dt_idt_j}\right]_{\substack{i\neq j \\ t_1 = t_2 = \dots = t_r = 0.}} = n(n-1)p_ip_j$$

We know, $Cov(x_ix_j) = E(x_ix_j) - E(x_i)E(x_j)$
$= n(n-1)p_ip_j - n^2p_ip_j = -np_ip_j$; $i\neq j$.

8.8 (a) Discrete Uniform or Rectangular Distribution

Among the discrete distributions, the discrete uniform distribution is the simplest one. A random variable x is said to have discrete uniform distribution if it assumes a finite set of values each with an equal probability of occurence. The probability function is given by

$$P(x) = \frac{1}{x}; \quad x = 1, 2 \dots\dots n \qquad \dots\dots\dots\dots(8.48)$$

If a fair die is tossed the possible out-comes are 1, 2, 3, 4, 5 and 6 each with probability $\frac{1}{6}$.

Hence in this case $P(x) = \frac{1}{6}$. Thus the probability is uniform for all values of the random variable x.

8.8 (b) Continuous Uniform or Rectangular Distribution

A random variable is said to have a continuous uniform distribution over the interval a to b if its probability density function (p. d. f.) is given by,

$$f(x) = \frac{1}{b-a}; \quad a \le x \le b,$$

Remarks

1) a and b are the two parameters of the distribution.
2) The graph of uniform p. d. f. f(x) is given below :

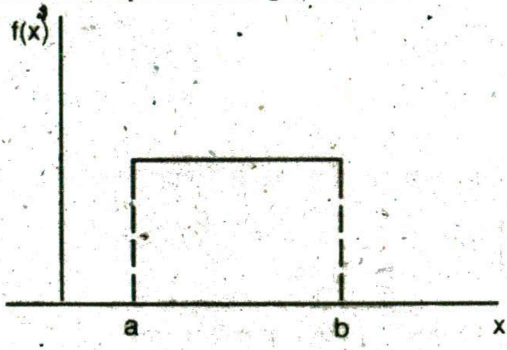


Fig. 8.1 Rectangular or uniform distribution.

Properties of Uniform Distribution :

Mean (μ) : $\mu = E(x) = \int_a^b x f(x)\, dx = \int_a^b x \frac{1}{b-a} dx = \frac{1}{2} \frac{b^2 - a^2}{(b - a)} = \frac{a+b}{2}$

Variance (σ^2) : We know, $\mu_2' = E(x^2) = \int_a^b x^2 f(x) dx$

$$= \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{3} \frac{b^3 - a^3}{(b-a)} = \frac{a^2 + ab + b^2}{3}$$

Now variance, $\sigma^2 = \mu'_2 - \mu'^2_1 = \frac{(b - a)^2}{12}$

We can easily show that, $E(x^r) = \frac{b^{r+1} - a^{r+1}}{(r + 1)(b - a)}$

It can be easily calculated that $\mu_3 = 0$ and $\mu_4 = \frac{(b - a)^4}{80}$

Therefore, $\beta_1 = 0$ and $\beta_2 = \frac{9}{5}$.

8.9 Normal Distribution

The most important and useful distribution in Statistics is the normal distribution. A random variable is said to have a normal distribution if its probability density function (p. d. f) is given by,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \; ; \; -\infty \le x \le \infty \qquad .....(8.49)$$

where μ and σ^2 are the mean and variance of the distribution.

Remarks :

1) μ and σ^2 are the two parameters of the distribution.
2) The normal variate is often expressed by $N(\mu,\sigma^2)$.
3) The assignment of probability is permissible,

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

$$= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}z^2} dz \qquad \left[\text{Putting } \frac{x-\mu}{\sigma} = z.\right]$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-t} \frac{dt}{\sqrt{2t}} \qquad \left[\text{Putting } \frac{z^2}{2} = t.\right]$$

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

$$= \frac{1}{\sqrt{\pi}} \Gamma\frac{1}{2} = 1, \text{ since } \Gamma\frac{1}{2} = \sqrt{\pi}.$$

4) The graph of f(x) is a famous bell shapped curve. The top of the bell is directly above the mean μ. For large values of σ, the curve tends to flaten out and for small values of σ, it has a sharp peak

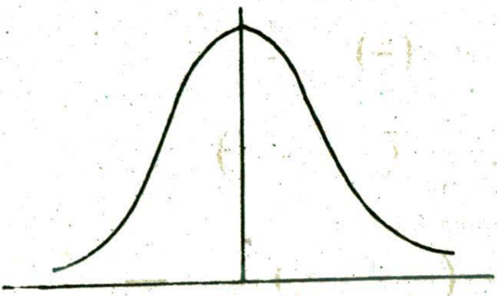


Fig. 8.2 Normal distribution.

Derivation of Normal Distribution (limiting form of Poisson Distribution) :

Normal distribution is a limiting form of the poisson distribution with the parameter $m \to \infty$ and $x \to \infty$

The probability function of the poisson distribution with parameter m is given by

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

The starling's approximation to x!, for large x is

$$x! = \sqrt{2\pi}\, e^{-x} x^{x+\frac{1}{2}}$$

Therefore,

$$\lim_{\substack{m\to\infty \\ x\to\infty}} p(x) = \lim_{m\to\infty} \frac{e^{-m} m^x}{\sqrt{2\pi} e^{-x} - x^{x+\frac{1}{2}}}$$

$$= \lim_{m\to\infty} \frac{e^{x-m}}{\sqrt{2\pi m}} \left(\frac{m}{x}\right)^{x+\frac{1}{2}}$$

$$= \frac{1}{\sqrt{2\pi m}} \lim_{m\to\infty} e^{x-m} \left(\frac{m}{x}\right)^{x+\frac{1}{2}}$$

Let $\frac{x-m}{\sqrt{m}} = z$ or, $x - m = z\sqrt{m}$ or, $x = m + z\sqrt{m}$

or, $\frac{x}{m} = 1 + \frac{z}{\sqrt{m}}$ or, $\frac{m}{x} = \left(1 + \frac{z}{\sqrt{m}}\right)^{-1}$

Again let $\Phi = e^{x-m} \left(\frac{m}{x}\right)^{x+\frac{1}{2}}$

$$= e^{z\sqrt{m}} \left(1 + \frac{z}{\sqrt{m}}\right)^{-\left(m + z\sqrt{m} + \frac{1}{2}\right)}$$

Taking logarithm we have,

$$\log \Phi = z\sqrt{m} - \left(m + z\sqrt{m} + \frac{1}{2}\right) \log\left(1 + \frac{z}{\sqrt{m}}\right)$$

$$= z\sqrt{m} - \left(m + z\sqrt{m} + \frac{1}{2}\right)\left(\frac{z}{\sqrt{m}} - \frac{z^2}{2m} + \frac{z^3}{3m\sqrt{m}} - \cdots\cdots\cdots\right)$$

$= z\sqrt{m} - z\sqrt{m} + \frac{z^2}{2} - z^2 +$ factors containing power of m in the denominator.

$$\therefore \lim_{m \to \infty} \log \Phi = -\frac{1}{2}z^2$$

$$\therefore \Phi = e^{-\frac{1}{2}z^2} = e^{-\frac{(x-m)^2}{2m}} \quad \text{(Putting the value of } z\text{)}$$

$$\text{Hence } \lim_{\substack{m \to \infty \\ x \to \infty}} p(x) = \frac{1}{\sqrt{2\pi m}} e^{-\frac{(x-m)^2}{2m}} \quad -\infty \le x \le \infty.$$

$$\therefore f(x) = \frac{1}{\sqrt{2\pi m}} e^{-\frac{(x-m)^2}{2m}} \quad -\infty \le x \le \infty$$

If we put $\frac{x-m}{\sqrt{m}} = \frac{x-\mu}{\sigma}$ (since mean and variance of poisson distribution are same)

We get finally,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \; ; -\infty \le x \le \infty.$$

This is the p. d. f. of the normal distribution with mean μ and variance σ^2.

Chief Characteristic of the Normal Distribution and Normal Probability Curve : The normal probability curve with mean μ and variance σ^2 is given by the equation

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \; ; -\infty \le x \le \infty.$$

and has the following properties.

1. The curve is bell shaped and symmetrical about the ordinate $x = \mu$.
2. As x increases numerically, f(x) decreases rapidly after the point $x = \mu$.
3. The maximum ordinate is at $x = \mu$ and is given by $y = \frac{1}{\sqrt{2\pi\sigma^2}}$

4. Points of inflexion are equidistant from the mean.
5. The curve extends to infinity on either side of the mean.
6. Arithmetic Mean, Median and Mode of the distribution coincide.
7. All odd moments are zero and $\beta_1 = 0, \beta_2 = 3$.
8. Linear combination of independent normal variates is also a normal variate.
9. Mean deviation about arithmetic mean is

$$\sqrt{\frac{2}{\pi}}\,\sigma = \frac{4}{5}\,\sigma \text{ (approx).}$$

10. Quartile deviation is equal to $\frac{2}{3}\sigma$.
11. Area property :

$P(\mu - \sigma \le x \le \mu + \sigma) = 0.6826.$

$P(\mu - 2\sigma \le x \le \mu + 2\sigma) = 0.9544.$

$P(\mu - 3\sigma \le x \le \mu + 3\sigma) = 0.9973.$

$P(-1.96 \le \frac{x-\mu}{\sigma} \le 1.96) = 0.95;$

$P(-2.58 \le \frac{x-\mu}{\sigma} \le 2.58) = 0.99.$

Mean and Other Moments of Normal Distribution :

Mean (μ) :

$$\mu = \mu_1' = E(x) = \int_{-\infty}^{\infty} x\, f(x)dx = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \int_{-\infty}^{\infty} (\mu+\sigma z) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \qquad \left[\text{Putting } z = \frac{x-\mu}{\sigma}\right]$$

$$= \mu + 0 = \mu$$

Since $z e^{-\frac{1}{2}z^2}$ is an odd function of z.

Odd order moments about mean :

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

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2r+1} e^{-\frac{1}{2}z^2} dz \left[\text{Putting } z = \frac{x-\mu}{\sigma}\right]$$

$$\text{or, } \mu_{2r+1} = \frac{\sigma^{2r+1}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2r+1} e^{-\frac{1}{2}z^2} dz = 0.$$

Since the integrand is an odd function of z.

Hence all odd order moments about mean are zero.

Even order moments about mean :

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

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2r} e^{-\frac{1}{2}z^2} dz \left[\text{Putting } z = \frac{x-\mu}{\sigma}\right]$$

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

Since the integrand is an even function of z, we have,

$$\mu_{2r} = \frac{2\sigma^{2r}}{\sqrt{2\pi}} \int_0^{\infty} (2t)^r e^{-t} \frac{dt}{\sqrt{2t}} \quad \left[\text{Putting } \frac{z^2}{2} = t\right]$$

$$= \frac{2^r \sigma^{2r}}{\sqrt{\pi}} \int_0^{\infty} e^{-t}\, t^{(r+\frac{1}{2})-1} dt.$$

$$= \frac{2^r \sigma^{2r}}{\sqrt{\pi}} \Gamma\left(r+\frac{1}{2}\right)$$

$$= \frac{2^r \sigma^{2r}}{\sqrt{\pi}} \left(r-\frac{1}{2}\right)\left(r-\frac{3}{2}\right)\left(r-\frac{5}{2}\right) \cdots \frac{3}{2}\frac{1}{2}\Gamma\frac{1}{2}$$

$$= \frac{1.3.5\ldots\ldots(2r-1)\,\sigma^{2r}\sqrt{\pi}}{\sqrt{\pi}}, \text{ Since } \Gamma\frac{1}{2} = \sqrt{\pi}$$

$$= 1.3.5\ldots\ldots(2r-1)\sigma^{2r}.$$

Therefore, when r=1 ; $\mu_2 = \sigma^2$ = Variance.

Again when r = 2 ; $\mu_4 = 1.3\sigma^4 = 3\sigma^4$.

and $\mu_3 = 0$, as we have obtained that odd order moments are zero.

Hence $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = 0$ and $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 3$.

These two values generally identify the type of the distribution.

Moment Generating Function of Normal Distribution :

The m. g. f of a normal variate about orgin is given by,

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

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

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(\mu + \sigma z)}\, e^{-\frac{1}{2}z^2}\, dz. \quad \left[\text{Putting, } z = \frac{x-\mu}{\sigma}\right]$$

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

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

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

$$= e^{\mu t + \frac{\sigma^2 t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\, du \quad [\text{Putting } z - \sigma t = u]$$

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

The moment generating function (m. g. f.) of normal distribution about mean is given by,

$M\mu(t) = E[\,e^{t(x-\mu)}\,] = e^{-\mu t}E(e^{tx})$

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

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

Hence,

$$M\mu(t) = 1 + \frac{t^2\sigma^2}{2} + \frac{\left(\frac{t^2\sigma^2}{2}\right)^2}{2!} + \frac{\left(\frac{t^2\sigma^2}{2}\right)^3}{3!} + \ldots\ldots + \frac{\left(\frac{t^2\sigma^2}{2}\right)^r}{r!} + \qquad \ldots\ldots(8.50)$$

Now the co-efficient of $\frac{t^r}{r!}$ gives μ_r. the rth moment about mean. Since there is no term with odd power of t, all moments of odd order about mean vanish, i, e. $\mu_{2r+1}=0$, which follows the earlier result.

And the even moments μ_{2r} = Co-efficient of $\frac{t^{2r}}{(2r)!}$ in (8.50) which is equal to

$$\frac{\sigma^{2r}(2r)!}{2^r\, r!}$$

$$= \frac{\sigma^{2r}\,[2r\,(2r-1)\,(2r-2)\ldots\ldots5,\,4,\,3,\,2,\,1]}{2^r\, r!}$$

$$= \frac{\sigma^{2r}\,[1.3.5\ldots\ldots..(2r-1)]\,[2.4.6\ldots\ldots(2r-2)2r]}{2^r\, r!}$$

$$= \frac{\sigma^{2r}\,[1.3.5\ldots\ldots..(2r-1)]\,2^r\,[1.2.3\ldots\ldots(r-1)r]}{2^r\, r!}$$

$= \sigma^{2r}\, 1.3.5\ldots..(2r-1)$, which is equivalent to the earlier result.

Standardised Normal Variate :

A variate is said to be a standardised normal variate if it is distributed normally with mean zero and variance unity.

Thus if, $x \sim N(\mu,\sigma^2)$, then $z = \frac{x-\mu}{\sigma}$ is a standardised normal variate with $E(z) = 0$ and var $(z) = 1$ and we write $z \sim N(0, 1)$.

The p.d. f is $f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2} \quad -\infty \le z \le \infty$ $\qquad \ldots\ldots.(8.51)$

Area Property of Normal Probability Integral :

If $x \sim N(\mu,\sigma^2)$, then the probability for the interval from the mean μ to the value x_1 is given by,

$$P(\mu \le x \le x_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu}^{x_1} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Let $\frac{x-\mu}{\sigma} = z$; $dx = \sigma dz$. when $x = \mu, z = 0$

and when $x = x_1$, $z = \frac{x_1 - \mu}{\sigma} = z_1$ (say)

$$\therefore P(\mu \le x \le x_1) = P(0 \le z \le z_1) = \frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{-\frac{1}{2}z^2} dz.$$

Where z is the standardised normal variate. The definite integral $\int_0^{z_1} f(z)dz$ is known as normal probability integral and the area under standard normal curve between the ordinate $z = 0$ and $z = z_1$. These areas have been tabulated for different value of z_1, at an interval of .01. [Such a table is provided by Biometrika Tables for Statistician Vol-1 by E. S. Pearson and O.H. Hartley P.P. 104-110.]

Example 8.5 A random variate x is normally distributed with mean 12 and standard deviation 4. Find out the probability of the following :

i) $x \ge 20$ ii) $x \le 20$ iii) $0 \le x \le 12$.

Solution : Here we have μ=12 and σ = 4.

i) when $x = 20$, $z = \frac{20-12}{4} = 2$

$\therefore P(x \ge 20) = P(z \ge 2) = P(0 \le z \le \infty) - P(0 \le z \le 2)$

$= 0.5 - 0.4772 = 0.0228.$

ii) $P(x \le 20) = P(z \le 2) = P(-\infty \le z \le 0) + P(0 \le z \le 2)$

$= 0.5 + 0.4772 = 0.9772.$

iii) $P(0 \le x \le 12) = P(-3 \le z \le 0) = P(0 \le z \le 3) = 0.49865.$

Importance of Normal Distribution in Statistics :

Normal distribution plays a very important role in Statistics because of the following reasons :

1) Most of the distributions occurring in practice e. g. Binomial, Poisson, Hyper-geometric distribution etc. can be approximated by the normal distribution under some assumptions. Moreover, many of the sampling distributions e. g. student's t, F and χ^2 tends to normality for large samples.

2) Even if the variable is not normally distributed, it can sometimes be brought to normal form by simple transformation of variable. For example, if the distribution of x is skewed, the distribution of $\sqrt{x}$ might come out to be normal.

3) The distribution has attractive mathematical properties which are very useful from theoretical point of view.

4) The proofs of all the tests of significance in sampling are based upon the fundamental assumption that the population from which the samples have been drawn is normal.

5) Normal distribution finds large application in statistical quality control theory.

Log Normal Distribution : The positive random variable x is said to have a log normal distribution if log x is normally distributed. The p. d. f. of x is given by

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}[\log x - \mu]^2} \quad ; x > 0. \qquad \text{....(8.52)}$$

Moments : The rth moment about origin is given by

$\mu_r' = E(x^r) = E(e^{ry})$, where $y = \log x$ or, $x = e^y$.

$= M_y(r)$, which is the m.g. f. of y, r being the parameter.

$= e^{\mu r + \frac{1}{2}r^2\sigma^2}$, since $y = \log x \sim N(\mu, \sigma^2)$.

Remarks :

1) For a particular case if we take $\mu = \log a$, $a > 0$.

then $\mu_r' = e^{r \log a + \frac{1}{2}r^2\sigma^2}$, $= a^r \; e^{\frac{1}{2}r^2\sigma^2}$.

Now taking r =1, $\mu_1' = a\, e^{\frac{\sigma^2}{2}}$

and if, r = 2, $\mu_2' = a^2 e^{2\sigma^2}$

$$\therefore \mu_2 = \mu_2' - \mu_1'^2 = a^2 e^{\sigma^2}\left(e^{\sigma^2} - 1\right)$$

2) Log normal distribution arises in problem of economics, biology, geology, and reliability theory. In particular, it arises in the study of dimension of particals under pulverization.

3) If $x_1, x_2 \ldots\ldots x_n$ is a set of independently identically distributed random variable such that mean of log x_i is μ and its variance is σ^2, then the product $x_1 x_2 \ldots\ldots x_n$ is asymptotically distributed according to log normal distribution with mean μ and variance $n\sigma^2$.

8.10 Gamma Distribution

A random variable is said to have a gamma distribution with parameter n if its probability density function is given by

$$f(x) = \frac{e^{-x}x^{n-1}}{\Gamma n}; \quad 0 \le x \le \infty, \; n > 0. \qquad \ldots(8.53)$$

and is denoted by G(n).

Remarks :

1) The function $\int_0^\infty e^{-x}x^{n-1}\,dx$ is knwon as gamma function and is denoted by Γn.

2) The assignment of probability is permissible, since

$$\int_0^\infty f(x)\,dx = \frac{1}{\Gamma n}\int_0^\infty e^{-x}x^{n-1}\,dx = \frac{1}{\Gamma n}\Gamma n = 1.$$

3) A continuous random variable having the following p. d. f. is said to have a gamma distribution with parameter λ and n if

$$f(x) = \frac{\lambda^n e^{-\lambda x}x^{n-1}}{\Gamma n}, \quad \begin{matrix} 0 \le x \le \infty \\ n, \lambda > 0 \end{matrix} \qquad \ldots\ldots(8.54)$$

and is denoted by $G(\lambda, n)$.

4) The cumulative distribution function (c.d.f) is called the Incomplete Gamma Function and is denoted by

$$F(p) = \frac{1}{\Gamma n}\int_0^p e^{-x}x^{n-1}\,dx; \quad \begin{matrix} x > 0 \\ n > 0 \end{matrix} \qquad \ldots\ldots\ldots(8.55)$$

Properties of Gamma Distribution :

Mean $(\mu) = \mu_1' = E(x) = \int_0^{\infty} x f(x)\, dx$

$$= \frac{1}{\Gamma n}\int_0^{\infty} x\, e^{-x} x^{n-1}\, dx = \frac{1}{\Gamma n}\int_0^{\infty} e^{-x} x^{n}\, dx.$$

$$= \frac{\Gamma(n+1)}{\Gamma n} = \frac{n\Gamma n}{\Gamma n} = n$$

Variance, (σ^2) :

We know, $\mu_2' = E(x^2) = \int_0^{\infty} x^2 f(x)\, dx.$

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

$$= \frac{1}{\Gamma n}\int_0^{\infty} e^{-x} x^{n+1}\, dx = \frac{\Gamma(n+2.)}{\Gamma n} = \frac{(n+1)\, n\Gamma n}{\Gamma n} = n(n+1).$$

$\therefore$ Variance, $\sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = n(n+1) - n^2 = n.$

Third moment (μ_3) :

We know, $\mu_3' = E(x^3) = \int_0^{\infty} x^3 f(x)\, dx$

$$= \frac{1}{\Gamma n}\int_0^{\infty} x^3 e^{-x} x^{n-1}\, dx = \frac{1}{\Gamma n}\int_0^{\infty} e^{-x} x^{n+2}\, dx$$

$$= \frac{\Gamma(n+3)}{\Gamma n} = \frac{(n+2)(n+1)\, n\Gamma n}{\Gamma n}$$

$= n(n+1)(n+2).$

$\therefore \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3. = n(n+1)(n+2) - 3n(n+1)n + 2n^3 = 2n.$

Fourth moment (μ_4) : We know, $\mu_4' = E(x^4) = \int_0^{\infty} x^4 f(x)dx.$

$$= \frac{1}{\Gamma n}\int_0^{\infty} x^4 e^{-x} x^{n-1}\, dx = \frac{1}{\Gamma n}\int_0^{\infty} e^{-x} x^{n+3}\, dx$$

$$= \frac{\Gamma(n+4)}{\Gamma n} = \frac{(n+3)(n+2)(n+1)\,n\,\Gamma n}{\Gamma n} = n(n+1)(n+2)(n+3).$$

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

$$= n(n+1)(n+2)(n+3) - 4n^2(n+1)(n+2) + 6n(n+1)n^2 - 3n^4$$

$= 3n^2+6n$ (on simplification)

Therefore, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(2n)^2}{n^3} = \frac{4}{n}$ and $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3n^2+6n}{n^2} = 3 + \frac{6}{n}$

The Moment Generating Function of Gamma Distribution :

The m. g. f. about origin of the gamma distribution is given by

$$M(t) = E(e^{tx}) = \frac{1}{\Gamma n}\int_0^{\infty} e^{tx} e^{-x} x^{n-1}\,dx$$

$$= \frac{1}{\Gamma n}\int_0^{\infty} e^{-x(1-t)} x^{n-1}\,dx$$

$$= \frac{1}{\Gamma n}\int_0^{\infty} e^{-z}\left(\frac{z}{1-t}\right)^{n-1}\frac{dz}{1-t} \quad [\text{Putting } z = x(1-t)]$$

$$= \frac{1}{\Gamma n\,(1-t)^n}\int_0^{\infty} e^{-z} z^{n-1}\,dz$$

$$= \frac{1}{(1-t)^n\,\Gamma n}\Gamma n = \frac{1}{(1-t)^n} = (1-t)^{-n} \qquad \ldots\ldots\ldots\ldots(8.56)$$

Differentiating M(t) once, twice etc. with respect to t and putting t = 0, we get the same result of the moments.

Remarks :

1) Like poisson distribution, mean and variance of gamma distribution are same.
2) As $n \to \infty$, $\beta_1 = 0$ and $\beta_2 = 3$. Hence the distribution tends to normal distribution as n becomes very large.
3) For more general gamma distribution,

$$dF(x) = \frac{\lambda^n e^{-\lambda x} x^{n-1}}{\Gamma n}\,dx, \qquad 0 \le x \le \infty, \quad \lambda, n > 0$$

The m. g. f. is given by, $M(t) = \left(1 - \frac{t}{\lambda}\right)^{-n}$ ....(8.57)

Theorem 8.1 The sum of two independent gamma variates with parameters m and n is also a gamma variate with parameter m + n.

Proof : Let x and y be two independent gamma variates with parameters m and n respectively. The m. g. f. of the sum $z = (x+y)$ is given by

$M_z(t) = M_{x+y}(t) = M_x(t)\, M_y(t).$

$=(1 - t)^{-m} (1 - t)^{-n} = (1 - t)^{-(m+n)},$

which is the m. g. f. of a gamma variate with parametrr m + n. Hence the result.

Remarks : This result can be generalised for any number of independent gamma variates.

8.11 Beta Distribution

Beta Distribution (First Kind) : A random variate is said to have a beta distribution of first kind if its probability density function is given by,

$$f(x) = \frac{1}{B(m,n)} x^{m-1} (1 - x)^{n-1}, \quad \begin{matrix} 0 \le x \le 1 \\ m,n > 0 \end{matrix} \qquad ....(8.58)$$

and is denoted by $B_1(m,n)$.

Remarks :

1) m and n are two parameters of the distribution.
2) The assignment of probability is permissible since,

$$\int_0^1 \frac{1}{B(m,n)} x^{m-1} (1 - x)^{n-1} dx$$

$$= \frac{1}{B(m,n)} \int_0^1 x^{m-1} (1 - x)^{n-1} dx = \frac{B(m,n)}{B(m,n)} = 1.$$

3) The cumulative distribution function is called the Incomplete Beta Function and is denoted by,

$$F(q) = \int_0^q \frac{1}{B(m,n)} x^{m-1} (1 - x)^{n-1} dx ; \qquad \begin{matrix} 0 \le x \le 1 \\ m,n > 0. \end{matrix}$$

Moments of Beta Distribution :

The rth moment about origin is given by,

$$\mu_r' = \int_0^1 x^r f(x)\,dx = \int_0^1 \frac{x^r x^{m-1}(1-x)^{n-1}}{B(m,n).}dx$$

$$= \frac{1}{B(m,n)} \int_0^1 x^{m+r-1}(1-x)^{n-1}dx = \frac{1}{B(m,n)} B(m+r,n).$$

$$= \frac{\Gamma(m+r)\Gamma n\,\Gamma(m+n)}{\Gamma(m+n+r)\Gamma m\,\Gamma n.} = \frac{\Gamma(m+r)\Gamma(m+n)}{\Gamma(m+n+r)\Gamma m}$$

In particular, when r=1,

$$\text{Mean} = \mu = \mu_1' = \frac{\Gamma(m+1)\Gamma(m+n)}{\Gamma(m+n+1)\Gamma m} = \frac{m\Gamma(m)\Gamma(m+n)}{(m+n)\Gamma(m+n)\Gamma m}$$

$$= \frac{m}{m+n}$$

$$\text{when } r=2\,;\ \mu_2' = \frac{\Gamma(m+2)\Gamma(m+n)}{\Gamma(m+n+2)\Gamma m} = \frac{(m+1)\,m\,\Gamma m\,\Gamma(m+n)}{(m+n+1)\,(m+n)\Gamma(m+n)\Gamma m}$$

$$= \frac{m(m+1)}{(m+n)(m+n+1)}$$

$$\therefore \text{Variance, } \sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = \frac{m(m+1)}{(m+n)(m+n+1)} - \frac{m^2}{(m+n)^2}$$

$$= \frac{mn}{(m+n)^2(m+n+1)} \quad \text{(on simplification).}$$

Similarly μ_3 and μ_4 can be obtained and the values of β_1 and β_2 can be calculated.

Beta Distribution (Second Kind) : A random variable is said to have a beta distribution of second kind if its probability density function is given by

$$f(x) = \frac{1}{B(m,n)} \frac{x^{m-1}}{(1+x)^{m+n}}, \quad \begin{cases} 0 \le x \le \infty \\ m,n > 0 \end{cases} \qquad \text{..........(8.59)}$$

and is denoted by $B_2(m,n)$.

If we put $1 + x = \frac{1}{y}$ in the above p. d. f. we get the beta distribution of the first kind,

$$f(y) = \frac{1}{B(m,n)} y^{m-1}(1-y)^{n-1}$$

If we put $x = \frac{1}{1+y}$ in the beta distribution of the first kind we get (8.59)

Beta Function : The function $\int_0^1 x^{m-1}(1-x)^{n-1}dx,\ 0 < x \leq 1$ is called the beta function and is denoted by B(m,n).

Relationship Between Beta and Gamma Function :

We know, $\Gamma m \Gamma n = \int_0^{\infty} e^{-x} x^{m-1}dx \int_0^{\infty} e^{-y} y^{n-1}dy.$

$$= \int_0^{\infty}\int_0^{\infty} e^{-(x+y)} x^{m-1} y^{n-1} dxdy.$$

Let $u = x + y$, $\quad v = \frac{x}{x+y}$

$\therefore x = uv,\ y = u(1-v)$ and $dxdy = |J|\,dudv$

$$\text{where } |J| = \begin{vmatrix} \frac{dx}{du} & \frac{dy}{du} \\ \frac{dx}{dv} & \frac{dy}{dv} \end{vmatrix} = u$$

As x and y range from 0 to ∞, u ranges from 0 to ∞ and v ranges from 0 to 1.

$$\therefore \Gamma m \Gamma n = \int_0^{\infty}\int_0^{1} e^{-u} (uv)^{m-1}\{u(1-v)\}^{n-1} u\,dudv.$$

$$= \int_0^{\infty} e^{-u} u^{m+n-1} du \int_0^{1} v^{m-1}(1-v)^{n-1} dv.$$

$$= \Gamma(m+n)\, B(m,n).$$

$$\therefore B(m,n) = \frac{\Gamma m \Gamma n}{\Gamma(m+n)}$$

Example 8.6 Find the value of $B\left(\frac{1}{2}, \frac{1}{2}\right)$ and hence $\Gamma\left(\frac{1}{2}\right)$.

Solution : We know, $\Gamma m \Gamma n = \int_0^{\infty} e^{-x} x^{m-1}dx \int_0^{\infty} e^{-y} y^{n-1}dy.$

$$= \int_0^{\infty}\int_0^{\infty} e^{-(x+y)} x^{m-1}.y^{n-1} dx\,dy. \quad .....(8.60)$$

Let us put, $x = r\,Cos^2\theta$; $y = r\,Sin^2\theta$.

$\therefore dx\,dy = |J|\,dr\,d\theta = 2r\cos\theta\sin\theta\,dr\,d\theta.$

As x and y range from 0 to ∞ ; r ranges from 0 to ∞ and θ ranges from 0 to $\frac{\pi}{2}$.

Therefore (8.60) becomes, $\Gamma m\,\Gamma n = 2\int_0^{\infty}\int_0^{\frac{\pi}{2}} e^{-r} r^{m+n-1}\cos^{2m-1}\theta\,\sin^{2n-1}\theta\,dr\,d\theta.$

$$= \int_0^{\infty} e^{-r} r^{m+n-1}\,dr\; 2\int_0^{\frac{\pi}{2}} \cos^{2m-1}\theta\,\sin^{2n-1}\theta\,d\theta.$$

$$= \Gamma(m+n)\; 2\int_0^{\frac{\pi}{2}} \cos^{2m-1}\theta\,\sin^{2n-1}\theta\,d\theta.$$

or, $\dfrac{\Gamma m\,\Gamma n}{\Gamma(m+n)} = 2\int_0^{\frac{\pi}{2}} \cos^{2m-1}\theta\,\sin^{2n-1}\theta\,d\theta.$

or, $B(m,n) = 2\int_0^{\frac{\pi}{2}} \cos^{2m-1}\theta\,\sin^{2n-1}\theta\,d\theta.$

Now, $B\left(\frac{1}{2},\frac{1}{2}\right) = 2\int_0^{\frac{\pi}{2}} d\theta = \pi$

Again, $B\left(\frac{1}{2},\frac{1}{2}\right) = \dfrac{\Gamma\frac{1}{2}\,\Gamma\frac{1}{2}}{\Gamma 1} = \pi,$ Since, $\Gamma 1 = 1$

or, $\left\{\Gamma\frac{1}{2}\right\}^2 = \pi$

$\therefore \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$

Example 8.7 If x and y are independent gamma variate with parameters m and n respectively then show that the variates $u = x + y$, $v = \frac{x}{x+y}$ are independent and that u is a $G(m+n)$ variate and v is a $B_1(m,n)$ variate.

Solution : We have, $f(x) = \frac{1}{\Gamma m} e^{-x} x^{m-1}, \quad \begin{cases} 0 \le x \le \infty \\ m > 0. \end{cases}$

and $f(y) = \frac{1}{\Gamma n} e^{-y} y^{n-1}, \quad \begin{cases} 0 \le y \le \infty \\ n > 0. \end{cases}$

Since x and y are independently distributed, their joint probability differential is given by,

$$dF(x,y) = f(x)\, f(y)\, dxdy = \frac{1}{\Gamma m\, \Gamma n} e^{-(x+y)} x^{m-1} y^{n-1} dxdy.$$

Now, $u = x + y$; $v = \frac{x}{x+y}$

$\therefore x = uv$; $y = u(1 - v)$. Then $dx\,dy = |J|\, du\, dv = u\, du\, dv$.

As x and y range from 0 to ∞ ; u ranges from 0 to ∞ and v ranges from 0 to. 1.

Hence the joint distribution of u and v is given by,

$$dF(u,v) = \frac{1}{\Gamma m \Gamma n} e^{-u} (uv)^{m-1} \{u(1-v)\}^{n-1} u\, du\, dv.$$

$$= \frac{1}{\Gamma m \Gamma n} e^{-u} u^{m+n-1} du\, v^{m-1} (1-v)^{n-1} dv.$$

$$= \frac{e^{-u} u^{m+n-1}}{\Gamma(m+n)} du. \frac{v^{m-1} (1-v)^{n-1}}{B(m,n)} dv.$$

This shows that u and v are idependently distributed as $G(m+n)$ and $B_1(m,n)$ variate respectively.

Example 8.8 If x and y are independent gamma variate with parameters m and n respectively ; show that

$u = x + y$ and $v = \frac{x}{y}$ are independent and that u is a $G(m+n)$ variate and v is a $B_2(m,n)$ variate.

Solution : As in Example (8.7) we have,

$$dF(x,y) = \frac{1}{\Gamma m \Gamma n} e^{-(x+y)} x^{m-1} y^{n-1} dxdy.$$

Since $u = x + y$ and $v = \frac{x}{y}$ we have, $x = \frac{uv}{1+v}, y = \frac{u}{1+v}$

and $dxdy = |J|\, du\, dv = \frac{u}{(1+v)^2} du\, dv.$

As x and y range from 0 to ∞, both u and v range from 0 to ∞. Therefore the joint probability distribution of u and v becomes,

$$dF(u,v) = \frac{1}{\Gamma m \Gamma n} e^{-u} \left(\frac{uv}{1+v}\right)^{m-1} \left(\frac{u}{1+v}\right)^{n-1} \frac{u}{(1+v)^2} du dv$$

$$= \frac{e^{-u} u^{m+n-1}}{\Gamma(m+n)} du \frac{1}{B(m,n)} \frac{v^{m-1}}{(1+v)^{m+n}} dv, \quad 0 \le u, v \le \infty ;$$

showing that u and v are independently distributed as $G(m+n)$ and $B_2(m,n)$ variate respectively.

Remarks : The above two examples lead to the following important results.

If x is a $G(m)$ variate and y is an independent $G(n)$ variate, then

1) $x+y$ is a $G(m+n)$ variate i.e. the sum of two independent gamma variates is also a gamma variate.

2) $\frac{x}{y}$ is a $B_2(m,n)$ variate i.e. the ratio of two independent gamma variates is a beta variate of second kind.

3) $\frac{x}{x+y}$ is a $B_1(m,n)$ variate.

8.12 Exponential Distribution

A random variable is said to have an exponential distribution with parameter $\lambda > 0$ if its p.d. f. is given by

$$f(x) = \lambda e^{-\lambda x}, \begin{cases} x \ge 0, \\ \lambda \ge 0 \end{cases} \qquad \text{............(8.61)}$$

The ordinate of the frequency curve is the highest at $x = 0$ and it decreases as x increases. The frequency curve of this distribution is shown in Fig 8.3.

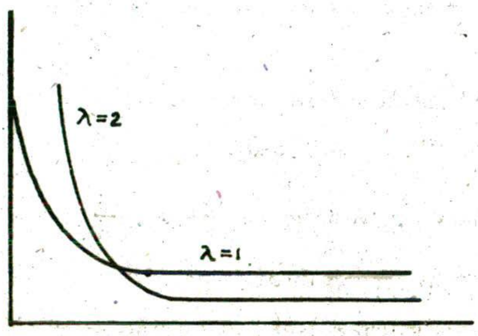


Fig. 8.3 Exponential distribution with $\lambda = 1$, $\lambda = 2$.

Properties of the Distribution :

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

$$= -xe^{-\lambda x}\Big]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Variance : We know variance $\sigma^2 = \mu_2 = \mu_2' - \mu_1'^2$

It can be easily shown that $\mu'_2 = E(x^2) = \frac{2}{\lambda^2}$.

$$\therefore \sigma^2 = \mu_2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Hence standard deviation $= \frac{1}{\lambda}$.

The Moment Generating Function of Exponential Distribution :

The m. g. f. of the distribution is $M(t) = E(e^{tx}) = \lambda \int_0^{\infty} e^{tx} e^{-\lambda x} dx$

$$= \lambda \int_0^{\infty} e^{-x(\lambda - t)} dx.$$

$$= \frac{\lambda}{(\lambda - t)} = \left(1 - \frac{t}{\lambda}\right)^{-1} = \sum_{r=0}^{\infty} \left(\frac{t}{\lambda}\right)^r.$$

We know, $\mu'_r = E(x^r)$ = Co-efficient of $\frac{t^r}{r!}$ in M(t), which is equal to $\frac{r!}{\lambda^r}$, r = 1, 2, 3..........

$$\therefore \mu'_1 = \frac{1}{\lambda};\ \mu'_2 = \frac{2}{\lambda^2} \text{ and so on.}$$

The third moment and fourth moment come out to be $\mu_3 = \frac{2}{\lambda^3}$ and $\mu_4 = \frac{9}{\lambda^4}$.

Therefore, $\beta_1 = 4$ and $\beta_2 = 9$ which are independent of λ.

Remarks :

1) The exponential variate is an special case of $G(\lambda, n)$ variate when $n = 1$.
2) The mean and standard deviation are equal.
3) The distribution is highly skewed.

8.13 Cauchy Distribution

A random variable x is said to have a standard cauchy distribution if its p. d. f. is given by

$$f(x) = \frac{1}{\pi(1 + x^2)} ; -\infty \leq x \leq \infty \qquad \text{......(8.62)}$$

In this case x is termed as standard cauchy variate.

In general, cauchy distribution with parameters λ and μ has the following p. d. f.

$$f(x) = \frac{\lambda}{\pi[\lambda^2 + (x - \mu)^2]} , \quad \begin{cases} \lambda > 0 \\ -\infty \leq x \leq \infty \end{cases} \qquad \text{.....(8.63)}$$

Characteristic Function of Cauchy Distribution :

The characteristic function of cauchy distribution is given by

$$\varphi(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} e^{itx} \frac{\lambda}{\lambda^2 + (x - \mu)^2} dx.$$

Let us put $\frac{x-\mu}{\lambda} = y \quad \therefore dx = \lambda dy.$

The range remain unchanged i. e. $-\infty \leq y \leq \infty$.

$$\text{Then, } \varphi(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} e^{it(\mu+\lambda y)} \cdot \frac{dy}{1+y^2}$$

$$= e^{it\mu} \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{it\lambda y}}{1+y^2} dy.$$

From the knowledge of Contour Integration we have,

$$\int_{-\infty}^{\infty} \frac{e^{it\lambda y}}{1+y^2} dy = \pi e^{-\lambda |t|}$$

Therefore the ch. function of the cauchy distribution becomes,

$$\varphi(t) = e^{it\mu} e^{-\lambda |t|}; \qquad \lambda > 0.$$

For standard cauchy distribution,

$$\varphi(t) = e^{it\mu} e^{-|t|}.$$

Additive Property of Cauchy Distribution : If x_1 and x_2 are independent cauchy variates with parameters (λ_1, μ_1) and (λ_2, μ_2) then x_1+x_2 is also a cauchy variate with parameters $(\lambda_1 + \lambda_2, \mu_1 + \mu_2)$.

Proof : $\phi_{x_1+x_2}(t) = \phi_{x_1}(t)\ \phi_{x_2}(t)$
(Since x_1 and x_2 are independent).
$=e^{it(\mu_1+\mu_2)-(\lambda_1+\lambda_2)|t|}$.

From the uniqueness theorem the result follows. This property can be extended for n independent cauchy variates.

Since $\varphi(t)$ in (8.63) does not exist at t = o, the mean of the cauchy distribution does not exist. Also the higher moments of cauchy distribution do not exist.

The arithmetic mean of a set of observations of cauchy distribution is also a cauchy distribution. In other words, in a cauchy distribution, the arithmetic mean of a sample of any size gives exactly as much information as a single variate x.

Moments of Cauchy Distribution :

$$E(x) = \int_{-\infty}^{\infty} x f(x)\, dx = \frac{\lambda}{\pi}\int_{-\infty}^{\infty}\frac{x}{\lambda^2+(x-\mu)^2}dx.$$

$$=\frac{\lambda}{\pi}\int_{-\infty}^{\infty}\frac{(x-\mu)+\mu}{\lambda^2+(x-\mu)^2}dx.$$

$$=\mu\frac{\lambda}{\pi}\int_{-\infty}^{\infty}\frac{dx}{\lambda^2+(x-\mu)^2}+\frac{\lambda}{\pi}\int_{-\infty}^{\infty}\frac{(x-\mu)}{\lambda^2+(x-\mu)^2}dx.$$

$$=\mu.1+\frac{\lambda}{\pi}\int_{-\infty}^{\infty}\frac{z}{\lambda^2+z^2}dz.$$

The integral $\int_{-\infty}^{\infty}\frac{z}{\lambda^2+z^2}dz$ is not completely convergent, its principal value,

viz. $\lim_{n\to\infty}\int_{-n}^{n}\frac{z}{\lambda^2+z^2}dz$ exists and is equal to zero.

Therefore, in general sense the mean of cauchy distribution does not exist. But if we assume that the mean of the cauchy distribution exists (by taking the principal value) then it is located at μ. Also, obviously, the probability curve is symmetrical about the point $x = \mu$, hence for this distribution the mean, median and mode coincide at the point $x = \mu$.

$$\text{Now, } \mu_2 = E(x - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$

$$\frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\lambda^2+(x-\mu)^2} dx,$$ which does not exist since the integral is not convergent.

Thus in general, for the cauchy distribution μ_r $(r \geq 2)$ do not exist.

8.14 Laplace Distribution

A continuous random varible x is said to have laplace distribution if the p. d. f. is given by,

$$f(x) = \frac{1}{2} e^{-|x|}, \quad -\infty \leq x \leq \infty, \qquad \text{.........(8.64)}$$

Characteristic function of laplace distribution is given by,

$$\varphi(t) = \frac{1}{2} \int_{-\infty}^{\infty} e^{itx} e^{-|x|} dx.$$

$$= \frac{1}{2}\left[\int_{-\infty}^{\infty} \text{Cos } tx\, e^{-|x|} dx + i \int_{-\infty}^{\infty} \text{Sin } tx\, e^{-|x|} dx \right].$$

$$= \frac{1}{2} \cdot 2 \int_{0}^{\infty} \text{Cos } tx\, e^{-|x|} dx.$$

Since the integrands in the first and second integrals are even and odd functions of x respectively.

$$\therefore \varphi(t) = \int_{-\infty}^{0} \text{Cos } tx\, e^{-x} dx.$$

$$= 1 - t^2 \int_{0}^{\infty} e^{-x} \text{Cos } tx\, dx \qquad \left[\begin{matrix}\text{On integation by} \\ \text{parts}\end{matrix}\right]$$

$$= 1 - t^2 \varphi(t)$$

$$\therefore \varphi(t) = \frac{1}{(1+t^2)}.$$

8.15 Pearsonian System of Frequency Curves

A set of frequency curves was developed in the first memoir of Karl Pearson in 1895 and in two subsequent papers in 1908 by assigning appropriate values of a, b_o, b_1 and b_2 in the following first order differential equation :

$$\frac{dy}{dx} = \frac{-y(x+a)}{b_0 + b_1x + b_2x^2} \qquad \text{.............(8.65)}$$

For obtaining the equation Karl Pearson considered the following characteristics :

1) A frequency distribution generally starts at zero, i. e. from a low frequency, rises to a maximum and again falls to the low frequency. Thus the frequency curve is generally unimodal. If the curve is represented by

$y = f(x)$, then $\frac{dy}{dx} = 0$ when $x = -a$.

2) At the ends of the frequency curves there is a high contact with the axis of x. i. e. $\frac{dy}{dx} = 0$ when $y = 0$.

3) The first four moments of the distribution are sufficient to determine the frequency curve.

Determination of the Constants of the Equation in Terms of Moments

Multiplying both sides of (8.65) by x^n and rearranging we get,

$$(b_0x^n + b_1x^{n+1} + b_2x^{n+2})\frac{dy}{dx}dx = -y(x^{n+1} + ax^n)\,dx$$

Integrating by parts over the entire range of the variate x.

We have, $$\left[(b_0x^n + b_1x^{n+1} + b_2x^{n+2})\,y\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\{nb_0x^{n-1} + (n+1)b_1x^n$$

$$+ (n+2)b_2x^{n+1}\}\,y\,dx = -\int_{-\infty}^{\infty}(x^{n+1} + ax^n)y\,dx.$$

Assuming the high contact at the extremities so that,

$[x^rf(x)]_{-\infty}^{\infty} = 0$ i. e. $x^rf(x) \to 0$ as $x \to \infty$ or $x \to -\infty$; and also

we know, $\int_{-\infty}^{\infty} x^nf(x)dx = \mu_n$, the nth moment.

Considering that x is measured from the mean we get,

$n b_0 \mu_{n-1} + (n+1) b_1 \mu_n + (n+2) b_2 \mu_{n+1} = \mu_{n+1} + a\mu_n.$

Putting n = 1, 2 and 3 using $\mu_0 = 1$ and $\mu_1 = 0$, we get,

$$\left.\begin{aligned} &b_1 = a \\ &b_0 + 3b_2\mu_2 = \mu_2 \\ &3b_1\mu_2 + 4b_2\mu_3 = \mu_3 + a\mu_2 \\ &3b_0\mu_2 + 4b_1\mu_3 + 5b_2\mu_4 = \mu_4 + a\mu_3 \end{aligned}\right\} \quad \ldots(8.66)$$

Solving (8.66) we get

$$b_0 = \frac{\sigma^2(4\beta_2 - 3\beta_1)}{2(5\beta_2 - 6\beta_1 - 9)},\ b_1 = \frac{\sigma\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} = a$$

$$b_2 = \frac{(2\beta_2 - 3\beta_1 - 6)}{2(5\beta_2 - 6\beta_1 - 9)}.$$

where $\mu_2 = \sigma^2$, $\beta_1 = \frac{\mu^2_3}{\mu^3_2}$ and $\beta_2 = \frac{\mu_4}{\mu_2^2}$

Putting the value of a, b_0, b_1 and b_2 in (8.65)

we have,

$$\frac{dy}{dx} = \frac{-y^2(5\beta_2 - 6\beta_1 - 9)x + \sigma\sqrt{\beta_1}(\beta_2 + 3)}{(2\beta_2 - 3\beta_1 - 6)x^2 + \sigma\sqrt{\beta_1}(\beta_2 + 3)\,x + \sigma^2(4\beta_2 - 3\beta_1)} \quad \ldots(8.67)$$

Method of Getting Different Types of Distributions :

The solution of the differential equation (8.65) depends mainly on the nature of the roots of the equation $b_0 + b_1x + b_2x^2 = 0$. The discriminant of the equation is $b_1^2 - 4b_0b_2$. Let us define a quantity $k = \frac{b_1^2}{4b_0b_2}$ on which the nature of various distributions will be determined.

Type 1 : Roots of $b_0+b_1x+b_2x^2 = 0$ are real, unequal and of opposite signs, so that $k < 0$.

Shifting the origin to the mode i. e. x = -a we have,

$$\frac{1}{y}\cdot\frac{dy}{dx} = \frac{x}{B(x + a_1)(x - a_2)}$$

$$= \frac{1}{B}\left[\frac{a_1}{(a_1 + a_2)}\cdot\frac{1}{(x + a_1)} + \frac{a_2}{(a_1 + a_2)}\cdot\frac{1}{(x - a_2)}\right]$$

$$= \frac{m_1}{(x - a_1)} + \frac{m_2}{(x - a_2)}$$

where $m_1 = \frac{a_1}{B(a_1 + a_2)}$ and $m_2 = \frac{a_2}{B(a_1 + a_2)}$

Now we have, $\frac{1}{y}dy = \left[\frac{m_1}{(x+a_1)} + \frac{m_2}{(x-a_2)}\right]dx.$

Integrating we get,

$\log y = m_1 \log(x+a_1) + m_2 \log(x-a_2) + \log C'_0$

or, $y = C'_0(x+a_1)^{m1}(x-a_2)^{m2}$

$$= C_0\left(1+\frac{x}{\alpha_1}\right)^{m_1}\left(1-\frac{x}{\alpha_2}\right)^{m_2}, \qquad -\alpha_1 \le x \le \alpha_2$$

where $\frac{m_1}{\alpha_1} = \frac{m_2}{\alpha_2}$ and C_0 is a constant.

Type VI : Roots of $b_0 + b_1x + b_2x^2 = 0$, are real, unequal and of same sign i. e. $k > 0$. Here also changing the origin to the mode, $x = -a$

we have, $\frac{1}{y}\cdot\frac{dy}{dx} = \frac{x}{B(x+a_1)(x+a_2)}$

In this Case, let the roots are be a_1 and a_2, so that,

$a_1 = -\alpha_1, a_2 = -\alpha_2, \alpha_1$ and $\alpha_2 > 0$

$\therefore \frac{m_1}{\alpha_1} = \frac{-m_2}{\alpha_2}$. (Vide Type 1)

The equation of the curve reduces to

$$y = C_0\left(1 - \frac{x}{\alpha_1}\right)^{m_1}\left(1+\frac{x}{\alpha_2}\right)^{-m_2}$$

which can be written as $y = C_0 x^{-m_2}(x+p)^{m_1}$; $\qquad -p \le x \le \infty$

Type IV : Roots of the equation $b_0 + b_1x + b_2x^2 = 0$ are imaginary, so that $0 \le k \le 1$.

We have, $\frac{1dy}{ydx} = \frac{-(x+a)}{b_0 + b_1x + b_2x^2}$

shifting the origin to $x = -a$ we have

$$\frac{1}{y}dy = \frac{-x}{b_2[(x+c)^2 + d^2]}dx$$

$$= \frac{(x+c) - c}{b_2[(x+c)^2 + d^2]}dx.$$

Integrating we get,

$$\log y = \log C'_0 - \frac{1}{2b_2}\log[(x+c)^2+d^2] - \frac{c}{b^2d^2}\tan^{-1}\frac{(x+c)}{d}$$

or, $$y = C'_0[(x+c)^2+d^2]^{-\frac{1}{2b_2}} e^{\left[\frac{-c}{b^2d^2}\tan^{-1}\frac{(x+c)}{d}\right]}$$

$$= C_0\left(1+\frac{x^2}{a^2}\right)^{-1} e^{-m\tan^{-1}\frac{x}{a}}; \quad 1, m > 0$$
$$-\propto \leq x \leq \propto.$$

Type III : One root of $b_0 + b_1x + b_2x^2 = 0$ is infinite $b_2 = 0 \;\; b_1 \neq 0$

$k \to \propto$

we have, $\frac{1}{y}\frac{dy}{dx} = -\frac{(x+a)}{b_0+b_1x}$

shifting the origin we get,

$$\frac{1}{y}dy = \frac{-xdx}{b_1(x+c)} = \left[-\frac{1}{b_1} + \frac{c}{b_1(x+c)}\right]dx$$

Integrating we get,

$$\log y = \log C_0 - \frac{x}{b_1} + \frac{c}{b_1}\log(x+c)$$

or $$y = C_0\left(1+\frac{x}{c}\right)^p e^{-\frac{px}{c}}; \quad -c \leq x \leq \propto$$

Type VII : Both the roots of $b_0+b_1x+b_2x^2 = 0$ are infinite i. e. $b_2 = 0 = b_1$, so that $k = 0$ we have,

$$\frac{1}{y}dy = -\frac{x+a}{b_0}dx$$

Integrating we have $\log y = \log C_0 - \frac{1}{2b_0}(x+a)^2$

or, $$y = C e^{-\frac{1}{2b_0}(x+a)^2} \qquad -\propto < x \leq \propto.$$

This curve is well known normal curve. This curve can also be obtained from (8.67) by putting $\beta_1 = 0$ and $\beta_2 = 3$ since in that case $b_2 = 0 = b_1$ and hence $k = 0$.

Type V : Roots of $b_0 + b_1x + b_2x^2 = 0$ are real, equal and of same sign so that $k = 1$.

We have, $\frac{1}{y}\cdot\frac{dy}{dx} = -\frac{(x+a)}{b_2(x+d)^2}$

$= -\frac{x}{b_2(x+c)^2}$ by proper choice of origin.

$= -\frac{1}{b_2}\left[\frac{1}{(x+c)} - \frac{c}{(x+c)^2}\right]$

or, $\frac{1}{y}dy = -\frac{1}{b_2}\left[\frac{1}{(x+c)} - \frac{c}{(x+c)^2}\right]dx$

Integrating we get, $\log y = \log C'_0 - \frac{1}{b_2}\log(x+c) - \frac{c}{b_2}(x+c)^{-1}$

or, $y = C'_0 (x+c)^{-\frac{1}{b_2}} e^{-\frac{c}{b_2(x+c)}}$

$= C_0 x^{-p} e^{-\frac{q}{x}}$; $0 \le x \le \infty$, $p, q > 0$.

Type II : Roots of $b_0 + b_1x + b_2x^2 = 0$ are real, equal but of opposite sign so that $k = 0$.

We have, $\frac{1}{y}\cdot\frac{dy}{dx} = \frac{(x+a)}{b_2(x-a_1)(x+a_1)} = \frac{(x+a)}{b_2(x^2-a_1^2)}$

or, $\frac{1}{y}dy = \frac{(x+a)}{b^2(x^2-a_1^2)}dx$.

Integrating we get,

$\log y = \log C'_0 + \frac{1}{2b_2}\log(x^2 - a_1^2)$.

or, $y = C'_0 (x^2 - a^2)^{\frac{1}{2}b^2}$

$= C'_0\left(1 - \frac{x^2}{a_1^2}\right)^m$, $-a_1 \le x \le a_1$

This curve is symmetrical with the mode at the origin. We can obtain this curve by putting $\beta_1 = 0$ and $\beta_2 < 3$ in (8.67)

Thus seven important different types of Pearsonian Curves are obtained.

The following table shows the name of distribution with density function, moment generating function, mean and variance.

Table-8.1

Sl. No.	Name of distribution	Density	Moment generating function	Mean	Variacne	Remarks
1.	Binomial	$\binom{n}{x} p^x q^{n-x}; x = 0, 1, 2.....n$	$(q+pe^t)^n$	np	npq	mean > variance
2.	Poisson	$\frac{e^{-m}m^x}{x!}, x = 0, 1, 2......\infty$	$e^{m(e^{t-1})}$	m	m	mean = variance
3.	Negative Binomial	$\binom{x+r-1}{r-1} p^r q^x, x = 0, 1....\infty$	$p^r(1-qe^t)^{-r}$	$\frac{rq}{p}$	$\frac{rq}{p^2}$	mean < variance
4.	Geometric	$pq^x, x = 0, 1, 2.....\infty$	$p(1-qe^t)^{-1}$	$\frac{q}{p}$	$\frac{q}{p^2}$	i) Putting r=1 in negative binomial distribution we get, geometric distribution ii) mean<variance
5.	Hypergeometric	$\frac{\binom{m}{x}\binom{n}{r-x}}{\binom{m+n}{r}}, \begin{matrix} x=0, 1, 2...r \\ r>0 \end{matrix}$	—	$\frac{mr}{m+n}$	$\frac{mnr(m+n-r)}{(m+n)^2(m+n-1)}$	—
6.	Multinomial	$\frac{n!}{x_1!x_2!.....x_r!} p_1^{x_1} p_2^{x_2} ...p_r^{x_r}$, $x_i = 0, 1, 2.......n_i$, $\Sigma n_i = n.$	$\left(p_1e^{t_1}+p_2e^{t_2}+.....+p_re^{t_r}\right)^n$	np_i	$np_i(1-p_i)$	$Cov(x_ix_j) = -np_ip_j.$

Sl. No.	Name of distribution	Density	Moment generating function	Mean	Variacne	Remarks
7.	Uniform (Discrete)	$\frac{1}{x}$; $x = 0, 1 \ldots\ldots n$,	*	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	—
8.	Uniform (Continuous)	$\frac{1}{b-a}$; $a \le x \le b$.	*	$\frac{(a+b)}{2}$,	$\frac{(a-b)^2}{12}$	—
9.	Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $-\infty \le x \le \infty$;	$e^{\mu t + \frac{t^2\sigma^2}{2}}$	μ	σ^2	$\beta_1 = 0$; $\beta_2 = 3$.
10.	Gamma i)	$\frac{1}{\Gamma n} e^{-x} x^{n-1}$; $0 \le x \le \infty$	$(1-t)^{-n}$	n	n	mean = variance.
	ii)	$\frac{1}{\Gamma n} \lambda^n e^{-\lambda x} x^{n-1}$; $0 \le x \infty$	$\left(1 - \frac{t}{\lambda}\right)^{-n}$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$	
11.	a) Beta (1st kind)	$\frac{1}{B(m,n)} x^{m-1} (1-x)^{n-1}$, $0 \le x \le 1$	*	$\frac{m}{m+n}$	$\frac{mn}{(m+n)^2 (m+n+1)}$.	
	b) Beta (2nd Kind)	$\frac{1}{B(m,n)} \frac{x^{m-1}}{(1+x)^{m+n}}$; $0 \le x \le \infty$				
13.	Exponential	$\lambda e^{-\lambda x}$; $0 \le x \le \infty$	$\left(1 - \frac{t}{\lambda}\right)^{-1}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	i) mean = standard deviation. ii) Special case of $G(\lambda, n)$ when n = 1.
14.	Cauchy	$\frac{a}{\pi\{a^2+(x-\mu)^2\}}$, $-\infty \le x \le \infty$	$\varphi(t) = e^{it\mu} e^{-\lvert t \rvert \frac{a}{2}}$	μ; does not exist.		Variance and higher moments donot exist.
15.	Laplace	$\frac{1}{2} e^{-\lvert x \rvert}$; $-\infty \le x \le \infty$	$\varphi(t) = \frac{1}{(1+t^2)}$ (ch. function)	0	2	It is easy to find out ch. function, $\varphi(t)$

9. CORRELATION AND REGRESSION

9.1 Bivariate Distribution

In earlier chapters, we mainly concentrate our attention to univariate distributions, i.e. the distributions involving one variable only. We may come across some situations in which each item of a series may have two or more variables. The distribution in which we consider two variables simultaniously for each item of the series is knwon as bivariate distribution. The distribution of heights and weights of a group of persons, the ages of husbands and wives of a number of couples etc. are the examples of bivariate distribution.

9.2 Correlation

In a bivariate distribution, there may exists correlation or co-variation between the variables. If the change in one variable effects a change in the other variable, the variables are said to be correlated. If the increase (decrease) in one variable results in the corresponing increase (decrease) in the others i. e. if the changes are in the same direction, the variables are positively correlated. For example, the heights and weights of a group of persons is positively correlated. If the increase (decrease) in one variable results in the corresponding derease (increase) in the other i. e. in this case the changes are in the opposite direction the variables are said to be negatively correlated. For example, the volume and pressure of a perfect gas is negatively correlated. If the changes do not depict any of the above two types., the variables are not correlated.

Scatter Diagram : The diagrammetic way of representing bivariate data is called scatter diagram. Thus for a bivariate distribution $(x_i, y_i)i=1, 2,....n$, the diagram of the dots obtained by the values of the variates x and y along the x-axis and y-axis respectively in the x, y-plane gives the scatter diagram. From a scatter diagram it can be evidently ascertained whether there is any correlation exists among the variates or not.

Correlation Co-efficient : We have already discussed that

$var(x) = \frac{1}{n}\sum^{n}(x_i - \overline{x})^2$, where $\overline{x}$ = mean of x_i gives the idea of variation among the values of the variable x, similarly

$var(y) = \frac{1}{n}\sum^{n}(y_i - \overline{y})^2$, where $\overline{y}$ = mean of y_i gives the variance of y. And

$Cov(x, y) = \frac{1}{n}\sum^{n}(x_i - \overline{x})(y_i - \overline{y})$ gives the co-variance between the variables x and y i. e. the simultanious variation of x and y. But co-variance is not independent of units of x and y. To make it a unit free measure Karl Pearson in 1890 defined correlation co-efficient between x and y as,

$$r_{xy} = \frac{Cov(x, y)}{\sqrt{var(x)\ var(y)}} = \frac{S.P.(x, y)}{\sqrt{S.S.(x).\ S.S.(y)}} = \frac{s_{xy}}{s_x s_y}$$

$$= \frac{\frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left\{\frac{1}{n}\sum(x_i - \overline{x})^2\right\}\left\{\frac{1}{n}\sum(y_i - \overline{y})^2\right\}}} \qquad \text{..........(9.1)}$$

Algebrically (9.1) reduces to

$$r_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left\{\sum^{n} x_i^2 - \frac{(\sum x_i)^2}{n}\right\}\left\{\sum^{n} y_i^2 - \frac{(\sum y_i)^2}{n}\right\}}} \qquad \text{......(9.2)}$$

(9.2) is usually considered as the working formula for calculating the correlation co-efficient between x and y. r_{xy} is sometimes called the product moment correlation co-efficient or total correlation co-efficient or co-efficient of correlation.

By symmetry it can be easily shown that $r_{xy} = r_{yx}$, r_{xy} is denoted sometimes simply by r.

Correlation Table : When the number of pairs of observations are large, it can be expressed in a tabular form known as correlation table or bivariate frequency distribution in which both the variables are classified one along the row and the other along the column. The value in a particular cell is the frequency of the pair lying in particular combination of class intervals.

Table-9.1

Correlation table of ages of husbands and wives of 53 couples.

	Age groups of wives (x)						
Age gr. of husbands (y)	15-25	25-35	35-45	45-55	55-65	65-75	Total
15—25	1	1	—	—	—	—	2
25—35	2	12	1	—	—	—	15
35—45	—	4	10	1	—	—	15
45—55	—	—	3	6	1	—	10
55—65	—	—	—	2	4	2	8
65—75	—	—	—	—	1	2	3
Total	3	17	14	9	6	4	53

Effect of change of origin and scale :

Let the origin and scale of x_i be changed and a new variate u_i is defined as

$u_i = \frac{x_i - a}{h}$, where a = origin and h = scale of the variate x_i and similarly,

$v_i = \frac{y_i - b}{k}$, where b = origin and k = scale of the variate y_i.

So that we have,

$x_i = hu_i + a$

or, $\overline{x} = h\overline{u} + a.$

and $y_i = kv_i + b$

or, $\overline{y} = k\overline{v} + b.$

Putting the values of x_i, $\overline{x}$, y_i and $\overline{y}$ in (9.1) we have,

$$r_{xy} = \frac{hk\Sigma(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{h^2k^2\{\Sigma(u_i - \overline{u})^2\}\{\Sigma(v_i - \overline{v})^2\}}}$$

$$\frac{hk}{\sqrt{k^2h^2}} r_{uv} = r_{uv} \qquad \text{..............(9.3)}$$

If h and k are both positive we have $r_{xy} = r_{uv}$, which indicates that correlation co-efficient is independent on changes of origin and scale. The method of this type of calculation is called short-cut method.

Limits of Correlation Co-efficient : The correlation co-efficient between x and y takes values from - 1 to + 1 i. e. $-1 \leq r_{xy} \leq 1$.

Let us consider

$$\left\{\frac{(x_i-\overline{x})}{s_x} \pm \frac{(y_i-y)}{s_y}\right\}^2 \geq 0$$

or, $$\frac{(x_i-\overline{x})^2}{s^2_x} + \frac{(y_i-\overline{y})^2}{s^2_y} \pm \frac{2(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y} \geq 0$$

Taking summation over the entire range of x_i and y_i we have

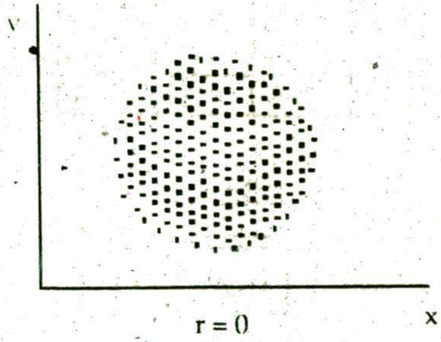
$$\frac{\Sigma(x_i-\overline{x})^2}{s^2_x} + \frac{\Sigma(y_i-\overline{y})^2}{s^2_y} \pm \frac{2\Sigma(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y} \geq 0$$

or, $$\frac{ns^2_x}{s^2_x} + \frac{ns^2_y}{s^2_y} \pm \frac{2ns_{xy}}{s_x s_y} \geq 0$$

or, $1 \pm r_{xy} \geq 0$, Since $\frac{s_{xy}}{s_x s_y} = r_{xy}$.

$\therefore\ -1 \leq r_{xy} \leq 1$. Hence proved.

Remark : Negative (Positive) value of r depends on the numerator i. e. the co-variance term. Different types of scatter diagrams for different values of r are given in Fig-9.1



r = 0

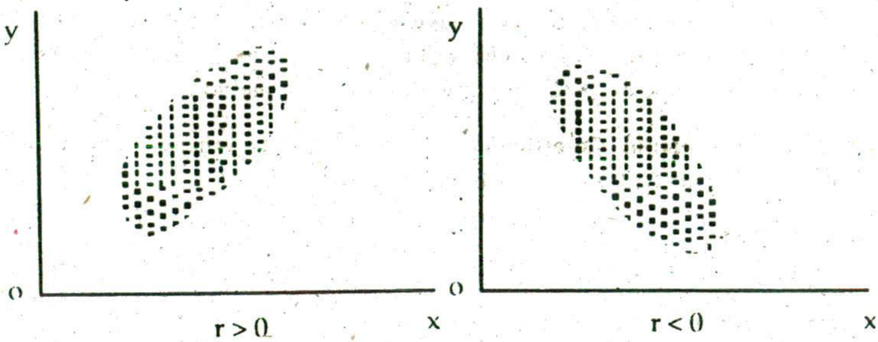


Fig. 9.1 Scatter diagrams for various values of r.

Example 9.1 Calculate the correlation co-efficient between the heights of father and son from the following data.

Height of father (in inches) :	65	66	67	68	69	70	71
Height of son (in inches) :	67	68	66	69	72	72	69

Solution : In the table-9.2 both the methods of calculation are shown.

Table-9.2

	Height of Father (x)	Height of Son (y)	x^2	y^2	xy	u= x-68	v= y-69	u^2	v^2	uv
	65	67	4225	4489	4355	-3	-2	9	4	6
	66	68	4356	4624	4488	-2	-1	4	1	2
	67	66	4489	4356	4422	-1	-3	1	9	3
	68	69	4624	4761	4692	0	0	0	0	0
	69	72	4761	5184	4968	1	3	1	9	3
	70	72	4900	5184	5040	2	3	4	9	6
	71	69	5041	4761	4899	3	0	9	0	0
Total	476	483	32396	33359	32864	0	0	28	32	20

From (9.2) we have,

$$r_{xy} = \frac{32864 - \dfrac{476 \times 483}{7}}{\sqrt{\left\{32396 - \dfrac{(476)^2}{7}\right\}\left\{33359 - \dfrac{(483)^2}{7}\right\}}}$$

$$= \frac{32864 - 32844}{\sqrt{(32396 - 32368)\,(33359 - 33329)}}$$

$$= \frac{20}{\sqrt{28 \times 30}} = 0.67 \text{ (app)} \qquad \therefore r_{uv} = \frac{nk}{\sqrt{k^2h^2}}$$

From (9.3) we have,

$$r_{uv} = \frac{20}{\sqrt{28 \times 30}} = 0.67 \text{ (app.)} \quad = \frac{\Sigma UV}{\sqrt{(\Sigma u^2)(\Sigma v^2)}} = \frac{\Sigma UV}{\sqrt{\Sigma U^2 \times \Sigma V^2}}$$

Therefore, it is shown numerically also that $r_{xy} = r_{uv}$.

Example 9.2 Calculate correlation co-efficient between the ages of husbands and wives given in Table-9.1.

Solution : We arrange the table as given in Table-9.1.

Table-9.3

Age of husbands (y)			Age of wives (x)									
			15-25	25-35	35-35	45-55	55-65	65-75				
		Mid Points	20	30	40	50	60	70	Total			
Age groups	Mid. Points	v \ u	-2	-1	0	1	2	3	f_y	vf_y	v^2f_v	uvf_v
15—25	20	-2	1	1	—	—	—	—	2	-4	8	6
25—35	30	-1	2	12	1	—	—	—	15	-15	15	16
35—45	40	0	—	4	10	1	—	—	15	0	0	0
45—55	50	1	—	—	3	6	1	—	10	10	10	8
55—65	60	2	—	—	—	2	4	2	8	16	32	32
65—75	70	3	—	—	—	—	1	2	3	9	27	24
	Total	f_u	3	17	14	9	6	4	53	16	92	86
		uf_u	-6	-17	0	9	12	12	10			
		u^2f_u	12	17	0	9	24	36	98			
		uvf_u	8	14	0	10	24	30	86			

Check

here $u = \frac{x - 40}{10}$ and $v = \frac{y - 40}{10}$.

Now $$r = \frac{\Sigma uvf_v - \frac{(\Sigma uf_u)(\Sigma vf_v)}{\Sigma f_u}}{\sqrt{\left\{\Sigma u^2 f_u - \frac{(\Sigma uf_u)^2}{\Sigma f_u}\right\}\left\{\Sigma v^2 f_v - \frac{(\Sigma vf_v)^2}{\Sigma f_v}\right\}}}$$

$$= \frac{86 - \frac{10 \times 16}{53}}{\sqrt{\left\{98 - \frac{(10)^2}{53}\right\}\left\{92 - \frac{(16)^2}{53}\right\}}}$$

$$= \frac{83}{\sqrt{(98-1.88)(92-4.83)}} = \frac{83}{\sqrt{96.12 \times 88.17}}$$

$= 0.912$. (app).

Example 9.3 If x and y are independent variables. Show that they are uncorrelated.

Solution : Since x and y are independent, we have

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

$= E(x - \overline{x})\, E(y - \overline{y}) = 0$

$\therefore r = 0$. Hence the result.

The converse of the result is not necessarily true i. e. variates may be uncorrelated but dependent. For this, an example of the following type may be considered, if x is a variate with a constant density function $f(x) = \frac{1}{2}, \; -1 \le x \le 1$ and if $y = x^2$.

then $E(x) = \int_{-1}^{1} x f(x)dx = \int_{-1}^{1} \frac{1}{2} x dx = 0$. So that $E(x)\,(E(y) = 0$.

Further more $E(xy) = E(x^3) = \int_{-1}^{1} \frac{1}{2} x^3 dx = 0$.

Hence $Cov(x, y) = E[(x - \overline{x})(y - \overline{y})] = E(xy) - E(x)\,E(y) = 0$

$\therefore r = 0$ i. e. x and y are uncorrelated.

However, for each value of x, there is only one possible value of y and for each value of y there are only two possible values of x. Therefore, x and y are far from being independent.

Example 9.4 x and y are two random variables with variances σ_x^2 and σ_y^2 $[\sigma_x^2 \neq 0 ; \sigma_y^2 \neq 0]$ respectively and r is the correlation co-efficient between them. If $u = x + ky$ and $v = x + \frac{\sigma_x}{\sigma_y} y$, find the value of k so that u and v are uncorrelated.

Solution : We know,

$u - E(u) = \{(x - E(x)\} + k\{y - E(y)\}.$

$v - E(v) = \{(x - E(x)\} + \frac{\sigma_x}{\sigma_y}\{y - E(y)\}.$

$Cov(u,v) = E[\{u - E(u)\} \{v - E(v)\}]$

$=E([\{x - E(x)\} + k\{y - E(y)\}] [\{x - E(x)\} + \frac{\sigma x}{\sigma y}\{y - E(y)\}])$

$=\sigma_x^2 + \frac{\sigma_x}{\sigma y} Cov (x, y) + k\, Cov (x, y) + k\frac{\sigma_x}{\sigma_y}\sigma_y^2$

$=\sigma_x^2 + \frac{\sigma_x}{\sigma y} r\, \sigma_x\sigma_y + kr\sigma_x\sigma_y + k\sigma_x\sigma_y$

$=\sigma_x^2 (1+r) + k\sigma_x\sigma_y(1+r)$

$=\sigma_x(1+r) (\sigma_x+k\sigma_y)$

u & v will be uncorrelated if $r_{uv} = 0$

$\therefore$ Cov (u,v) = 0.

That is, $\sigma_x(1+r) (\sigma_x+k\sigma_y) = 0$.

$\therefore \sigma_x+k\sigma_y = 0$ Since $\sigma_x \neq 0$ and $r \neq -1$.

or, $k = -\frac{\sigma_x}{\sigma_y}$.

Thus the value of k is determined.

Example 9.5 Let $y = -\frac{ax+c}{b}$. Prove that correlation co-efficient between x and y is - 1 if signs of a and b are alike and + 1 if they are different.

Solution : We know, $y = -\frac{ax + c}{b}$ or, $\overline{y} = -\frac{a\overline{x} + c}{b}$.

Thus $\overline{x} = -\frac{b\overline{y} + c}{a}$.

We have, $r = \dfrac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\{\Sigma(x_i - \overline{x})^2\}\{\Sigma(y_i - \overline{y})^2\}}}$.

$$= \frac{-\frac{a}{b}\Sigma(x_i - \overline{x})^2}{\pm\frac{a}{b}\Sigma(x_i - \overline{x})^2}$$

This means that $r = +1$ if the signs of a and b are different,

and $r = -1$ if they have the same sign.

Hence the result.

Example 9.6 If x and y are two correlated variables with the same standard deviation, say s and the correlation co-efficient, r.

Show that the correlation co-efficient between x and x+y is $\sqrt{\frac{(1+r)}{2.}}$

Solution: Let $u = x + y$ then $\overline{u} = \overline{x} + \overline{y}$.

$v(u) = v(x+y) = v(x) + v(y) + 2\ \text{Cov.}\ (x,y)$.

$= s^2 + s^2 + 2s^2 r$.

$= 2s^2(1+r)$.

$\text{Cov}\ (u,x) = E[(u - \overline{u})(x - \overline{x})]$

$= E[\{(x - \overline{x}) + (y - \overline{y})\}(x - \overline{x})]$

$= E(x - \overline{x})^2 + E(x - \overline{x})(y - \overline{y})$

$= s^2 + \text{Cov}\ (xy) = s^2 + s^2 r = s^2(1+r)$.

Therefore, the correlation co-efficient between u and x is

$$r_{ux} = \frac{s^2(1+r)}{\sqrt{s^2 s^2 2(1+r)}} = \frac{s^2(1+r)}{s^2\sqrt{2(1+r)}} = \sqrt{\frac{(1+r)}{2}}.$$ Hence proved.

Example 9.7 If x and y are uncorrelated, find the correlation co-efficient between $u = x + y$ and $v = x - y$.

Solution: Let $u = x + y$ or, $\overline{u} = \overline{x} + \overline{y}$

and $v = x - v$, or, $\overline{v} = \overline{x} - \overline{v}$

Now, $\text{Cov}(u,v) = E\{(u - \overline{u})(v - \overline{v})\}$

$= E[\{(x - \overline{x}) + (y - \overline{y})\}\{x - \overline{x}) - (y - \overline{y})\}]$

$= E\{(x - \overline{x})^2 - (y - \overline{y})^2\}$

$= E(x - \overline{x})^2 - E(y - \overline{y})^2$

$= s^2_x - s_y^2.$

where s_x^2 and s_y^2 are the variances of x and y respectively.

Now $v(u) = v(x+y) = v(x) + v(y) + 2\,\text{Cov}(x, y)$

$= v(x) + v(y)$. Since x and y are uncorrelated.

$= s^2_x + s^2_y.$

similarly $v(v) = s^2_x + s^2_y.$

Hence, $r_{uv} = \dfrac{\text{Cov}(u, v)}{\sqrt{v(u)\,v(v)}} = \dfrac{s^2_x - s^2_y}{s^2_x + s^2_y}.$

9.3 Regression

Correlation indicates whether there is any relation between the variables and correlation co-efficient measures the extent of relationship between them, whereas the regression measures the probable movement of one variable in term of the other. Therefore, regression is used for prediction problem.

The term "regression" was used by a famous Biometrician Sir. F. Galton (1822-1911) in connection with the inheritance of stature. But now it is widely used in Statistics.

Regression Lines : Let us consider that there exists association between x and y. In the scatter diagram for a particular value of x represented in the x-axis, we may consider a large number of observations along y-axis. We get a regression curve if we draw the x values and the corresponding mean values of y and the relationship is said to be expressed by means of curvilinear regression. If the curve is straight, it is called the line of regression and the regression is said to be linear, otherwise it is called curvilinear.

The line of regression is the straightline which gives the best fit to the bivariate frequency distribution in the least square sense. If the straight line be so chosen that the sum of square of the deviations parallel to the y-axis is minimum, we get a regression line of y on x and it gives the best estimate of y for any given value of x. On the other hand, if the sum of squares of the deviations parallel to the x-axis is minimum, the regression line of x on y is obtained and it gives the best estimate of x for any given value of y.

Let us suppose that (x_i, y_i) i=1, 2,.......n be a random sample from a bivariate distribution, y is dependent and x is independent variable. Let the regression line of y on x be

$$y = a + bx \qquad \text{.............(9.4)}$$

Following the principle of least squares method, the estimates of a and b can be obtained as below :

The observation y_i follows the model

$$y_i = a + bx_i + e_i \qquad \text{.............(9.5)}$$

where a is the intercept and b is the slope usually called the regression co-efficient of y on x and e_i's are random error componenets which are independently and normally distributed with 0 mean and variance σ^2.

From (9.5) we have,

$$e_i = y_i - a - bx_i$$

or, $$\sum_i e^2_i = s = \sum_i (y_i - a - bx_i)^2$$

Now, $$\frac{\delta s}{\delta a} = 0 \Rightarrow \sum_i y_i = na + b\sum_i x_i \qquad \text{............(9.6)}$$

and $$\frac{\delta s}{\delta b} = 0 \Rightarrow \sum_i x_i y_i = a\sum_i x_i + b\sum x^2_i \qquad \text{.........(9.6)}$$

These two equations are known as normal equations.

Considering (9.6) and (9.7) and dividing by n we get

$$a + b\,\overline{x} = \overline{y} \qquad \text{.............(9.8)}$$

$$a\,\overline{x} + \frac{b\sum x^2_i}{n} = \frac{\sum x_i y_i}{n} \qquad \text{............(9.9)}$$

Multiplying (9.8) by $\overline{x}$, we get the normal equations as

$$a\,\overline{x} + b\,\overline{x}^2 = \overline{x}\;\overline{y}$$

$$a\,\overline{x} + \frac{b\sum_i x_i^2}{n} = \frac{\sum_i x_i y_i}{n}$$

Subtracting we get,

$$b\left(\sum_i \frac{x_i^2}{n} - \overline{x}^2\right) = \sum_i \frac{x_i y_i}{n} - \overline{x}\,\overline{y}$$

$$\text{or, } \hat{b} = \frac{\sum_i x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}} = \frac{SP(x, y)}{SS(x).} = \frac{\frac{1}{n}\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\Sigma(x_i - \overline{x})^2} \quad .....(9.10)$$

Putting the value of b in (9.8) we have,

$$\hat{a} = \overline{y} - \frac{SP(x, y)}{SS(x)}\overline{x} \quad ....(9.11)$$

Thus the estimated values of a and b in (9.4) are obtained.

Therefore, the least square regression line of y on x in terms of value of a and b is, $y = \left(\overline{y} - \frac{SP(x, y)}{SS(x)}\overline{x}\right) + \frac{SP(x, y)}{SS(x)}x.$

$$\text{or, } (y - \overline{y}) = \frac{SP(x, y)}{SS(x)}(x - \overline{x})$$

$$\text{or, } (y - \overline{y}) = \frac{rs_y}{s_x}(x - \overline{x}) \quad ....(9.12)$$

Now considering the regression line of x on y as $x = a' + b'y$ and preceeding as above we have,

$$\hat{a}' = \overline{x} - \frac{SP(x, y)}{SS(y)}\overline{y} \text{ and } \hat{b}' = \frac{SP(x, y)}{SS(y)} = \frac{rs_x}{s_y}$$

Thus the least square regression line of x on y is

$$(x - \overline{x}) = \frac{SP(x, y)}{SS(y)}(y - \overline{y}) \text{ or, } (x - \overline{x}) = \frac{rs_x}{s_y}(y - \overline{y}) \quad ....(9.13).$$

Properties of the Regression Co-efficient :

a) Regression co-efficients are independent on change of origin but not of scale.

Let $u_i = \frac{x_i - a}{h}$, where a is origin and h is the scale of x_i

and $vi = \frac{y_i - b}{k}$, where b is origin and k is scale of y_i.

We have, $x_i = hu_i + a$ or, $\overline{x} = h\overline{u} + a$,

similarly, $y_i = kv_i + b$ or, $\overline{y} = k\overline{v} + b$.

and let us denote. $b_{y/x}$ as regression co-efficient of y on x and $b_{v/u}$ as regression co-efficient of v on u.

Now putting the value of x_i, y_i, $\overline{x}$ and $\overline{y}$ in (9.10) we have

$$b_{y/x} = \frac{hk\,\Sigma(u_i - \overline{u})(v_i - \overline{v})}{h^2\,\Sigma(u_i - \overline{u}^2)} = \frac{k}{h} b_{y/u}.$$

Proceeding in the above way we get,

$b_{x/y} = \frac{h}{k} b_{u/v}$, which shows that the regression co-efficients are independent on change of origin but not of scale.

(b) Correlation co-efficient is the geometric mean of the regression co-efficients.

We know, $b_{y/x} = \frac{SP(x, y)}{SS(x)} = \frac{rs_y}{s_x}$.

and also $b_{x/y} = \frac{SP(x, y)}{SS(y)} = \frac{rs_x}{s_y}$.

Now, $b_{y/x} \times b_{x/y} = r^2$. Therefore, $r = \pm\sqrt{b_{y/x} \times b_{x/y}}$ ....(9.14)

Hence proved.

Remarks :

1) We have $r = \frac{SP(x, y)}{\sqrt{SS(x)SS(y)}}$, $b_{y/x} = \frac{SP(x,y)}{SS(x)}$ and $b_{x/y} = \frac{SP(x, y)}{SS(y)}$.

Therefore, the sign of correlation co-efficient is the same as that of regression co-efficients because the sign of each of them depends on SP(x, y). Thus the sign of correlation co-efficient, r in (9.14) depends on regression co-efficients i. e. if the regression co-efficients are positive r is positive and if the regression co-efficients are negative r is negative.

2) If one of the regression co-efficients is greater than unity, the other must be less than unity.

Let us suppose that, one of the regression co-efficients say, $b_{y/x}$ is greater than unity i. e. $b_{y/x} > 1$ which implies that $\frac{1}{b_{y/x}} < 1$.

Also $r^2 \le 1$

$\therefore b_{y/x} \times b_{x/y} \le 1$

Hence $b_{y/x} \le \frac{1}{b_{x/y}} \le 1$. Hence proved.

c) Arithmetic mean of the regression co-efficients is greater than the correlation co-efficient.

We know that, Arithmetic mean $\ge$ Geometric mean

Therefore, $\frac{1}{2}(b_{y/x} + b_{x/y}) \ge \sqrt{b_{y/x} \times b_{x/y}}$.

or, $\frac{1}{2}(b_{y/x} + b_{x/y}) \ge r$. Hence proved.

Aliter : We have to show that

$$\frac{1}{2}(b_{y/x} + b_{x/y}) \ge r$$

or, $\frac{1}{2}\left(\frac{rs_y}{s_x} + \frac{rs_x}{s_y}\right) \ge r$

or, $\frac{s^2_y + s^2_x}{s_x s_y} \ge 2$

or, $(s^2_x + s^2_y - 2s_x s_y) \ge 0$

or, $(s_x - s_y)^2 \ge 0$

which is always true since the square of real quantity is greater than or equal to zero.

d) **Angle between two lines of regression :**

Equations of the lines of regression of y on x and that of x on y are-

$y - \overline{y} = \frac{rs_y}{s_x}(x - \overline{x})$ and $x - \overline{x} = \frac{rs_x}{s_y}(y - \overline{y})$ respectively.

Slopes of the lines are $\frac{rs_y}{s_x}$ and $\frac{s_y}{rs_x}$ respectively.

Let us consider that θ be the angle between the two regression lines then,

for acute angle,

$$\tan\theta = \frac{\frac{s_y}{rs_x} - \frac{rs_y}{s_x}}{1 + \frac{rs_y\, s_y}{s_x\, rs_x}}$$

$$= \frac{\frac{s_y(1-r^2)}{rs_x}}{\frac{r(s^2x + s^2y)}{rs^2x}} = \frac{1-r^2}{r}\left(\frac{s_x s_y}{s^2x + s^2_y}\right)$$

$$\therefore \theta = \tan^{-1}\left\{\frac{1-r^2}{r}\left(\frac{s_x s_y}{s^2x + s^2_y}\right)\right\}.$$

for obtuse angle,

$$\therefore \theta = \tan^{-1}\left\{\frac{r^2-1}{r}\left(\frac{s_x s_y}{s^2_x + s^2_y}\right)\right\}.$$

Case (i) $r = 0$ $\tan\theta = \infty$ $\therefore \theta = \frac{\pi}{2}$

Thus if two variables are uncorrelated, the lines of regression becomes perpendicular to each other.

Case (ii) If $r = \pm 1$ $\tan\theta = 0$ $\therefore \theta = 0$ or π.

In this case, the two regression lines are either coincide or they are parallel to each other but since the regression lines pass through the points $(\overline{x}, \overline{y})$ they cannot be parallel. Hence for perfect correlation positive or negative, the two regression lines coincide.

Example 9.8 Obtain the equations of the regression lines from the data given in Example 9.1 Also estimate of x for y = 70.

Solution : The equation of the regression line of y on x is

$$y - \overline{y} = b_{y/x}(x - \overline{x})\,; \quad \text{where, } b_{y/x} = \frac{SP(x,y)}{SS(x)} = \frac{20}{28} = 0.71 \text{ (app).}$$

Thus the regression equation of y on x becomes

$y - 69 = 0.71(x - 68)$; Since $\overline{x} = 68$ and $\overline{y} = 69$;

or, $y = 0.71x + 20.72$.

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

Where, $b_{x/y} = \frac{SP(xy)}{SS(y)} = \frac{20}{30} = 0.67$ (app).

Therefore the regression line is, $x - 68 = 0.67(y - 69)$.

or, $x = 0.67y + 21.77$.

The estimate of x for given y = 70 is given by $\hat{x} = 68.67$.

9.4 Rank Correlation

In some situations it is difficult to measure the values of the variables from bivariate distribution numerically, but they can be ranked. The correlation co-efficient between these two rank is usually called rank correlation co-efficient, given by Spearman (1904).

Let (x_i, y_i) ; i = 1, 2,......n, denote the ranks of the ith individual of two characteristics A and B respectively. Assuming that no two individuals are awarded the same rank in either classification, each of the variables x and y takes the values 1, 2,........n.

$$\text{Hence } \bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + \ldots\ldots + n) = \frac{n(n+1)}{2n} = \frac{(n+1)}{2}$$

$$s^2_x = \frac{1}{n}\sum_i(x_i - \bar{x}^2) = \frac{1}{n}\sum x_i^2 - \bar{x}^2.$$

$$= \frac{1}{n}[1^2 + 2^2 + \ldots\ldots + n^2] - \left(\frac{n+1}{2}\right)^2 = \frac{1}{n}\frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \frac{(n+1)}{2}\left[\frac{2n+1}{3} - \frac{n+1}{2}\right] = \frac{(n+1)(n-1)}{12} = \frac{n^2-1}{12} = s^2_y.$$

Let $d_i = x_i - y_i = (x_i - \bar{x}) - (y_i - \bar{y})$.

Squaring and summing over the range of i from 1 to n we get,

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

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

Dividing both sides by n we have

$$\left(\frac{\sum_i d_i^2}{n}\right) = s^2_x + s^2_y - 2s_{xy} = s^2_x + s^2_x - 2rs^2_x. \qquad \text{Since, } s^2_x = s^2_y.$$

$$= 2s^2_{\ x}(1 - r) = 2.\frac{n^2 - 1}{12}(1 - r)$$

$$r = 1 - \frac{6\sum di^2}{n(n^2 - 1)} \qquad ....(9.15)$$

Remark :

1) If $x_i = y_i$, $i = 1, 2......n$, all the di's reduces to zero and $r = + 1$.

2) If the ranking are as follows :

$x = 1, \quad 2, \quad 3, \quad ....n$

$y = n, \quad (n-1), \quad (n-2), \quad ....1.$

Then $r = - 1$.

Proof : Let us consider one case particularly when n is odd.

Let $n = 2m + 1$ then the di's are

$2m, 2m - 2, 2m - 4, ...2, 0,........2.....4,......(2m - 2), ...2m.$

$\therefore \sum d_i^2 = 2\{(2m)^2 + (2m-2)^2 + ... + 4^2 + 2^2\}$

$= 8\{m^2 + (m-1)^2 + +.....2^2+1^2\}.$

$$= \frac{8m(m+1)(2m+1)}{6}.$$

We know, for n, $r = 1 - \frac{6\sum d^2i}{n(n^2 - 1)}$. Putting $n = 2m + 1$

$$r = 1 - \frac{8m(m + 1)(2m + 1)}{(2m + 1)\{(2m + 1)^2 - 1\}}$$

$$= 1 - \frac{8m(m+1)}{4m^2 + 4m} = 1 - \frac{8m(m + 1)}{4m(m+1)} = -1.$$

In the same way it can be easily shown that for $n = 2m$, the result also follows.

3) We always have $\Sigma d_i = \Sigma(x_i - y_i) = n\overline{x} - n\overline{y} = 0$. This serves as a check on the calculations.

Example 9.9 The ranks of ten students in Mathematics and Statistics are as follows. Find the rank correlation co-efficient.

Mathematics :	3,	5,	8,	4,	7,	10,	2,	1,	6,	9.
Statistics :	6,	4,	9,	8,	1,	2,	3,	10,	5,	7.

Solution :

Table for calculation of rank correlation co-efficient

Table-9.4

Rank in Math. (x)	Rank in Stat (y)	$d_i = (x_i - y_i)$ differences	d_i^2
3	6	-3	9
5	4	1	1
8	9	-1	1
4	8	-4	16
7	1	6	36
10	2	8	64
2	3	-1	1
1	10	-9	81
6	5	1	1
9	7	2	4
Total		$\Sigma d_i = 0$	214

Rank correlation co - efficient, $r = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = -0.3$ (app).

Tied Ranks : When there is more than one item with the same value which are then said to be tied in the series, then the formula for calculating rank correlation co-efficient breaks down. Since in this case, each of the variable x and y does not assume the values 1, 2, 3,.....n and cosequently $\overline{x} \neq \overline{y}$. In that case, the most common method is to allocate to each member the mean of the ranks which the tied members would have if they were ordered. This is called the mid-rank method. As a result of this, following correction is made in the rank correlation co-efficient formula.

In the formula, we add the factor $\frac{m(m^2-1)}{12}$ to $\sum_i d_i^2$, where m is the number of items an item is repeated or tied. This correction factor is to be added for each tied value.

Example 9.10 Obtain the rank correlation co-efficient for the following data.

A :	68,	64,	75,	50,	64,	80,	75,	40,	55,	64
B :	62,	58,	68,	45,	81,	60,	68,	48,	50,	70

Solution :

Table for calculation of rank correlation co-efficient

Table-9.5

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

In the series A, the correction is to be applied twice, once for the value 75 which occurs twice (m = 2) and that for the value 64 which occurs thrice (m = 3). The total correction for series A is $\frac{2(2^2-1)}{12}+\frac{3(3^2-1)}{12}=\frac{5}{2}$.

Similarly, the correction for series B is $\frac{2(2^2-1)}{12}=\frac{1}{2}$ as the value 68 occurs twice.

$$\text{Thus, } r = 1 - \frac{6\left[\sum_i^n d_i^2 + \frac{5}{2}+\frac{1}{2}\right]}{n(n^2-1)} = 1 - \frac{6(72+3)}{10 \times 99} = 0.545 \text{ (app).}$$

9.5 Bivariate Normal Distribution

Two normally correlated continuous variables x and y are said to have bivariate normal distribution if their joint probability density function is given by

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \text{Exp.}\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right\}\right]$$

$-\infty \le x \le \infty$, and $-\infty \le y \le \infty$ ....(9.16)

Where μ_1 and σ_1^2 are the mean and variance of x, μ_2 and σ_2^2 are the mean and variance of y, and ρ is the correlation co-efficient between x and y.

The frequency surface representing a bivariate normal distribution is shown in Figure 9.2.

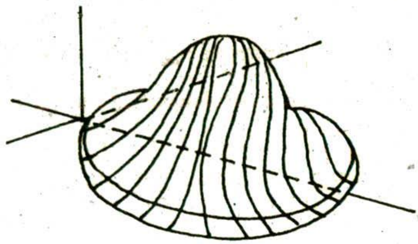


Fig. 9.2 Bivariate Normal Surface.

Moment Generating Function of Bivariate Normal Distribution : The moment generating function (m.g.f.) of bivariate normal distribution about the means μ_1 and μ_2 is given by

$$M(t_1, t_2) = E[\text{Exp}\{t_1(x-\mu_1) + t_2(y-\mu_2)\}]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \text{Exp}\{t_1(x-\mu_1) + t_2(y-\mu_2)\} f(x,y)\, dx\, dy.$$

which reduces to $\text{Exp}[\frac{1}{2}(t_1^2\sigma_1^2 + 2t_1t_2\rho\sigma_1\sigma_2 + t_2^2\sigma_2^2)]$

Now μ_{rs} = co-efficient of $\frac{t_1^r t_2^s}{r!s!}$ in the expansion of $M(t_1,t_2)$ where the first suffix corresponds to x and second suffix corresponds to y variate.

Marginal Density : Marginal density of x of the bivariate normal distribution is given by $\int_{-\infty}^{\infty} f(x,y).\,dy$, putting the value of f(x,y) and after simplification we get, $g(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma_1^2}}$; $-\infty \le x \le \infty$.

similarly, $h(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu_2)^2}{\sigma_2^2}}$; $-\infty \le y \le \infty$.

Hence it is seen that x is normally distributed with mean μ_1 and variance σ_1^2; y is also normally distributed with mean μ_2 and variance σ_2^2. If $\rho = 0$, $f(x,y) = g(x).h(y)$, which means that x and y are independetly normally distributed, but the zero correlation does not imply independence in general.

Conditional Density : The conditional density of x for given value of y is given by $f(x/y) = \frac{f(x,y)}{h(y)}$, which after simplification reduces to

$$\frac{1}{\sigma_1\sqrt{2\pi(1-\rho^2)}} e^{-\frac{1}{2\sigma_1^2(1-\rho^2)}\left[x - \left\{\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right\}\right]^2} \quad ; -\infty \le x \le \infty.$$

It is as like as univariate normal distribution with mean $\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2)$ and variance $\sigma_1^2(1 - \rho^2)$. Similarly we can show that the conditional distribution of y for given x is normal with mean $\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$ and variance $\sigma_2^2(1 - \rho^2)$.

9.6 Correlation Ratio

Correlation ratio, η is the appropriate measure of curvilinear relationship between the two variables. When the relationship is linear, the extent of association is measured by correlation co-efficient r. Therefore r measures the concentration of points about the straight line of the best fit and η measures the concentration of points about the curve of the best fit. $\eta = r$ if the regression is linear otherwise $\eta > r$ (see equation 9.21).

Let x_i, (i = 1, 2,......m) be the values of the variable x and its corresponding values of the variable y be y_{ij} with respective frequencies f_{ij}, (j = 1, 2.....n).

x / y	1	2	....	i	....	....m	Total
1	f_{11}	f_{21}	....	....f_{i1}	....	....f_{m1}	
|	|	|	|	|	|	|	
2	f_{12}	f_{22}	....	....f_{i2}	....	....f_{m2}	
|	|	|	|	|	|	|	
j	f_{1j}	f_{2j}	....	....f_{ij}	....	....f_{mj}	
|	|	|	|	|	|	|	
n	f_{1n}	f_{2n}	....	....f_{in}	....	....f_{mn}	
Total $\sum_j f_{ij}$	n_1	n_2	....	....n_i	.....	....n_m	$\sum_i n_i = n$
$T_i = \sum_j f_{ij}y_{ij}$	T_1	T_2	....	....T_i	....	....T_m	$\sum_i T_i = T.$

Though all the x's in the ith vertical array have the same value, the y's are different. The ith pair of values in the array is (x_i, y_j) with frequency f_{ij}. The first suffix 'i' indicates the vertical array and the second suffix 'j' indicates the position of y in the array.

Let $\sum_j f_{ij} = n_i$; $\sum_i\sum_j f_{ij} = \sum_i n_i = N$, say.

If $\overline{y_i}$ and $\overline{y}$ denote the means of the ith array and the weighted mean of all the array means, the weight being the frequency respectively, then

$$\overline{y_i} = \frac{\sum f_{ij}y_{ij}}{\sum_j f_{ij}} = \frac{\sum f_{ij}y_{ij}}{n_i} = \frac{T_i}{n_i} \text{ and } \overline{y} = \frac{\sum\sum f_{ij}y_{ij}}{\sum_i\sum_j f_{ij}} = \frac{\sum n_i \overline{y_i}}{\sum n_i} = \frac{T}{N}.$$

The correlation ratio of y on x, usually denoted by η_{yx} is given by the formula $\eta^2_{yx} = 1 - \frac{s^2_e}{s^2}$ ....(9.17)

where $s^2_e = \frac{1}{N}\sum\sum f_{ij}(y_{ij} - \overline{y_i})^2$ and $s^2 = \frac{1}{N}\sum\sum f_{ij}(y_{ij} - \overline{y})^2$.

$\sum\sum f_{ij}(y_{ij} - \overline{y})^2$ can be partitioned into two parts and a convenient expression can be developed as below :

$$Ns^2 = \sum\sum f_{ij}(y_{ij} - \overline{y})^2 = \sum\sum f_{ij}\{(y_{ij} - \overline{y_i}) + (\overline{y_i} - \overline{y})\}^2.$$

$$= \sum\sum f_{ij}(y_{ij} - \overline{y_i})^2 + \sum\sum f_{ij}(\overline{y_i} - \overline{y})^2 + 2\sum\sum f_{ij}(y_{ij} - \overline{y_i})(\overline{y_i} - \overline{y}).$$

The product term venishes due to $\sum_i f_{ij} (\overline{y_{ij}} - \overline{y_i}) = 0$

Therefore, $Ns^2 = \Sigma\Sigma f_{ij}(y_{ij} - \overline{y_i})^2 + \Sigma n_i(\overline{y_i} - \overline{y})^2$.

or, $Ns^2 = Ns^2_e + Ns^2_m$, where $s^2_e = \frac{1}{N}\Sigma\Sigma f_{ij}(\overline{y}_{ij} - \overline{y}_i)^2$,

and $s^2_m = \frac{1}{N}\Sigma n_i(\overline{y_i} - \overline{y})^2$.

$\therefore 1 - \frac{s^2_e}{s^2} = \frac{s^2_m}{s^2}$.

Now comparing (9.17) we have $\eta^2_{yx} = \frac{s^2_m}{s^2}$ ....(9.18)

The calculation of s^2_m can be done conveniently as follows :

$$Ns^2_m = \Sigma n_i(\overline{y_i} - \overline{y})^2 = \Sigma n_i \overline{y_i}^2 - N\overline{y}^2 = \sum \frac{Ti^2}{n_i} - \frac{T^2}{n}.$$

Remarks :

1) Since s^2_e and s^2 are non negative
$1 - \eta^2_{yx} \geq 0 \quad \therefore \eta^2_{yx} \leq 1$. and it follows that
$0 \leq \eta^2_{yx} \leq 1$. ....(9.19)

2) Since the sum of squares of the deviations in any array is minimum when measured from its means we have,

$$\Sigma\Sigma f_{ij}(y_{ij} - \overline{y_i})^2 \leq \Sigma\Sigma f_{ij}(y_{ij} - \hat{y}_{ij})^2 \qquad ....(9.20)$$

Where $\hat{y}_{ij}$ is the estimate of y_{ij} for given $x = x_i$ as given by the line of regression of y on x i.e. $\hat{y}_{ij} = a + bx_i$ $(j = 1, 2.....n)$.

But $\Sigma\Sigma f_{ij}(y_{ij} - \overline{y_i})^2 = Ns^2_e = Ns^2(1 - \eta^2_{yx})$.

and $\Sigma\Sigma f_{ij}(y_{ij} - \hat{y}_{ij})^2 = \Sigma\Sigma f_{ij}(y_{ij} - a - b_{xi})^2 = Ns^2(1 - r^2)$

Therefore, from the inequality given in (9.20) we have

$1 - \eta^2_{yx} \leq 1 - r^2$ i.e. $\eta^2_{yx} \geq r^2$

or, $|\eta_{yx}| \geq |r|$ ....(9.21)

Combining (9.19) and (9.21) we have, $0 \leq r^2 \leq \eta^2_{yx} \leq 1$.

Thus it can be concluded that the absolute value of the correlation ratio can never be smaller than the correlation co-efficient. When the regression of y on x is linear, the means of the array will be on the line of regression and we have $\eta^2_{yx} = r^2$. Thus $\eta^2_{yx} - r^2$ gives the departure of linearity of regression. If $\eta^2_{yx} = 1, s^2e = 0$

$\therefore \Sigma\Sigma f_{ij}(y_{ij} - \overline{y}_i)^2 = 0$ $\quad\therefore y_{ij} = \overline{y}_i$ for all $j = 1, 2.....n$; indicating that all the points are on the mean of the array. If the array means of y are closer to the grand mean, $\overline{y}$, η^2_{yx} approaches to zero.

3) $r_{xy} = r_{yx}$ but $\eta_{yx} \neq \eta_{xy}$.

4) Like correlation co-efficient, correlation ratio is independent of change of scale and origin.

Example 9.11 Find the correlation ratio of y on x from the data given in Example 9.2.

Solution : We are to calculate $\overline{y}$, s^2_y and s^2_{my}.

$$\overline{y} = 40 + \frac{10 \times 16}{53} = 43.02.$$

$$s_y = 10\sqrt{\frac{\Sigma v^2 fv}{\Sigma fv} - \left(\frac{\Sigma vfv}{\Sigma fv}\right)^2} = 10\sqrt{\frac{92}{53} - \left(\frac{16}{53}\right)^2}$$

$= 10\sqrt{1.6447} = 10 \times 1.28 = 12.8$ (app).

The Table 9.6 shows the calculation of s^2_{my}.

Table-9.6

Mean of cols ($\overline{y}$)	$f_{\overline{y}}$	$u' = \overline{y} - 41$	$f_{\overline{y}}u'$	$f_{\overline{y}}u'^2$
26.67	3	-14.33	-42.99	616.0467
31.76	17	-9.24	-157.08	1451.4192
41.34	14	0.43	6.02	2.5886
51.11	9	10.11	90.99	919.9089
60.00	6	19.00	114.00	2166.0000
65.00	4	24.00	96.00	2304.0000
Total	53		106.94	7459.9634

$$s^2_{my} = \frac{7459.9634}{63} - \left(\frac{106.94}{53}\right)^2 = 140.7540 - 4.0713 = 136.6827.$$

Therefore, $s_{my} = 11.69$, Now $\eta_{yx} = \frac{11.69}{12.8} = 0.913$ (app).

Remark : From the same data given in Example 9.2 and Example 9.11 we have shown that $|\eta_{yx}| \geq |r|$.

9.7 Intracless Correlation

Intraclass correlation means within class correlation. In biological and agricultural study one may often be interested to know how the members of a family or group are correlated among themselves with respect to some one of the same characteristics. For example, the correlation between the heights or weights of brothers in one or more families or the between yields of certain crop of one or more experimental blocks will give intraclass correlation.

Suppose we have k families $A_1, A_2, \ldots\ldots A_k$ with $n_1, n_2, \ldots\ldots n_k$ numbers of each and the measurements x_{ij}, $(i = 1, 2\ldots\ldots k ; j = 1, 2\ldots\ldots n_i)$ of the characteristic can be arranged as below :

$$\begin{array}{cccccc}
x_{11} & x_{21} & \cdots & x_{i1} & \cdots & x_{k1} \\
x_{12} & x_{22} & \cdots & x_{i2} & \cdots & x_{k2} \\
| & | & | & | & | & \\
x_{1j} & x_{2j} & \cdots & x_{ij} & \cdots & x_{kj} \\
| & | & | & | & | & | \\
x_{1n_1} & x_{2n_2} & \cdots & x_{in_i} & \cdots & x_{kn_k}
\end{array}$$

We shall have $n_i(n_i - 1)$ pairs (x_{ij}, x_{il}) $j \neq 1$ of observations in the ith family or group. There will be $\sum_{i=1}^{k} n_i \; (n_i - 1) = N$ say, pairs for all the k families or groups. If we prepare a correlation table, there will be $n_i(n_i - 1)$ entries for the ith group. The table will be symmetrical about the principal diagonal, x_{i1} occurs $(n_i - 1)$ times, x_{i2} occurs $(n_i - 1)$ times and hence for all the k families, we have $\sum_i (n_i - 1) \sum_j x_{ij}$ as the marginal total.

$\therefore \; \overline{x} = \overline{y} = \frac{1}{N}\{\Sigma(n_i - 1)\, \Sigma x_{ij}\}$ Similarly, $s^2_x = s^2_y = \frac{1}{N}\{\sum_i (n_i - 1)\, \Sigma(x_{ij} - \overline{x})^2\}$

Further $\text{cov}(x,y) = \sum_i \{\sum_{\substack{j\ l \\ j \neq l}} (x_{ij} - \overline{x})(x_{il} - \overline{x})\}$.

$$= \frac{1}{N} \sum_j \{\sum_{i=1}^{n_j} \sum_{l=1}^{n_i} (x_{ij} - \overline{x})(x_{il} - \overline{x}) - \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2\}.$$

$$= \frac{1}{N} [\sum_i n_i (\overline{x}_i - \overline{x}) n_i (\overline{x}_i - \overline{x}) - \sum_i \sum_j (x_{ij} - \overline{x})^2]$$

$$= \frac{1}{N} \sum n_i^2 (\overline{x}_i - \overline{x})^2 - \sum \sum (x_{ij} - \overline{x})^2]$$

Therefore, the intraclass correlation co-efficient is given by

$$r_{(xy)} = \frac{\text{Cov}(x,y)}{\sqrt{\text{var}(x)\ \text{var}(y)}} = \frac{\sum n_i^2 (\overline{x}_i - \overline{x})^2 - \sum\sum (x_{ij} - \overline{x})^2}{\sum\sum (n_i - 1)(x_{ij} - \overline{x}^2)} \quad .....(9.22)$$

If $n_i = n$ i.e. all the families or groups have the equal number of members, then

$$r_{(xy)} = \frac{n^2 \sum (\overline{x}_i - \overline{x})^2 - \sum\sum (x_{ij} - \overline{x})^2}{(n-1) \sum\sum (x_{ij} - \overline{x})^2}$$

$$= \frac{kn^2 s^2_m - kns^2}{kn(n-1)s^2}$$

$$= \frac{1}{(n-1)} \left\{ \frac{ns^2_m}{s^2} - 1 \right\} \quad ....(9.23)$$

where s^2 indicates the variance of x and s^2_m is the variance of the mean of the families.

From (9.21) we have, $1 + (n-1) r_{(xy)} = \frac{ns^2_m}{s^2} \geq 0$

$$\therefore r_{(xy)} \leq -\frac{1}{(n-1)} \quad ....(9.24)$$

since $\frac{s^2_m}{s^2} \leq 1$, $\quad 1 + (n-1) r_{(xy)} \leq n$

$$\therefore r_{(xy)} \leq 1. \quad ....(9.25)$$

Now combining (9.24) and (9.25) we have the range of $r_{(xy)}$ as

$$-\frac{1}{(n-1)} \leq r_{(xy)} \leq 1.$$

Example 9.12 In five families of 3, the heights of brothers are in inches as below :

Families

		1	2	3	4	5
Brothers	1	69	70	71	72	73
	2	70	71	72	73	74
	3	71	72	73	74	75

Find intraclass correlation co-efficient.

Solution : Here k = 5, n = 3, N = 15.

$\overline{x} = 72,\ \overline{x}_1 = 70,\ \overline{x}_2 = 71,\ \overline{x}_3 = 72,\ \overline{x}_4 = 73,\ \overline{x}_5 = 74.$

$$s^2_m = \frac{1}{5}\Sigma(\overline{x_i} - \overline{x})^2$$

$$s^2 m = \frac{1}{5}[4+1+0+1+4] = \frac{10}{5} = 2.$$

$$s^2 = \frac{1}{kn}\Sigma\Sigma(x_{ij} - \overline{x})^2 = \frac{30}{15} = 2.$$

Threfore, the intraclass co-efficient, $r_{(xy)} = \frac{1}{(n-1)}\left\{\frac{ns^2_m}{s^2} - 1\right\}$.

$$= \frac{1}{2}\left\{\frac{3 \times 2}{2} - 1\right\} = 1.$$

9.8 Multiple and Partial Correlation and Regression

We have already discussed the correlation between two variables only. But often, it is necessary to obtain the correlation between three or more variables. If one variable is influenced by the combined effect of group of other variables we get multiple correlation and multiple regression. On the other hand, if one variable is influenced by another variable eliminating the linear effect of the other variables we get partial correlation and partial regression.

For example, the yield of a crop/acre (x_1) may be ifluenced by soil fertility (x_2), amount of rainfall (x_3), types of irrigations (x_4) and so on. Now if we are interested to acertain the association between x_1 and the combined effect of x_2, x_2, x_4 and so on we get multiple correlation and the degree of association is known as multiple correlation co-efficient and is denoted by $R_{1.234\ldots\ldots}$

Again if we are interested to ascertain the association between x_1 and x_2 when the linear effect of $x_3, x_4, \ldots\ldots$etc. are eliminated, we get partial correlation and the degree of association is given by partial correlation co-efficient, denoted by $r_{12\cdot34\ldots\ldots}$.

Regression Plane and Determination of Regression Co-efficient : For simplicity sake we consider 3 variables x_1, x_2 and x_3 only. The equation of the regression plane of x_1 on x_2 and x_3 is given by

$$x_1 = b_{12\cdot3}x_2 + b_{13\cdot2}x_3 \qquad \ldots\ldots\ldots\ldots(9.26)$$

assuming that the variables are measured from their respective means, b's are usually called the partial regression co-efficients which can be estimated by the least square method. The normal equations can be written as, $\sum x_2(x_1 - b_{12\cdot3}x_2 - b_{13\cdot2}x_3) = 0$

$\sum x_3(x_1 - b_{12\cdot3}x_2 - b_{13\cdot2}\, x_3) = 0.$

Expressing the equations in terms of standard deviations and correlation co-efficients we have,

$$r_{13}s_1 = b_{12.3}s_2 + b_{13.2}\, r_{23}s_3$$
$$r_{13}s_1 = b_{12.3}\, r_{23}s_2 + b_{13.2}s_3 \qquad \ldots\ldots\ldots\ldots(9.27)$$

where r_{ij} is the correlation co-efficient between x_i and x_j and s_i is the standard deviation of x_i.

Solving (9.27) we have,

$$b_{12.3} = -\frac{s_1}{s_2}\frac{r_{12} - r_{23}\cdot r_{13}}{(1 - r_{23}^2)} = -\frac{s_1}{s_2}\frac{\Delta_{12}}{\Delta_{11}} \qquad \ldots\ldots\ldots(9.28)$$

$$\text{similarly, } b_{13.2} = -\frac{s_1}{s_3}\frac{\Delta_{13}}{\Delta_{11}} \qquad \ldots\ldots\ldots\ldots(9.29)$$

where Δ_{ij} is the co-factor of the element in the ith row and jth column in the determinent $\Delta = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$ in which $r_{ij} = r_{ji}$.

Substituting the value of $b_{12.3}$ and $b_{13.2}$ in (9.26) we have

$$x_1 = -\frac{s_1}{s_2}\frac{\Delta_{12}}{\Delta_{11}}x_2 - \frac{s_1}{s_3}\frac{\Delta_{13}}{\Delta_{11}}x_3 \qquad \ldots\ldots\ldots\ldots(9.30)$$

or, $\frac{\Delta_{11}}{s_1} x_1 + \frac{\Delta_{12}}{s_2} x_2 + \frac{\Delta_{13}}{s_3} x_3 = 0$

The residual of second order $x_{1.23}$ is defined by

$x_1 - b_{12.3} x_2 - b_{13.2} x_3$.

Remark : In general, the equation of the regression plane of x_1 on x_2, x_3, x_4 etc. is,

$x_1 = b_{12.34 \ldots n} x_2 + b_{13.34 \ldots n} x_3 + \ldots\ldots + b_{1n.23 \ldots (n-1)} x_n$.

where $b_{12.34 \ldots n} = -\frac{s_1}{s_2} \frac{\Delta_{12}}{\Delta_{11}}$,

$b_{13.24 \ldots n} = -\frac{s_1}{s_3} \frac{\Delta_{13}}{\Delta_{11}}$ and similarly

$b_{1n.23 \ldots (n-1)} = -\frac{s_1}{s_n} \frac{\Delta_{1n}}{\Delta_{11}}$

where Δ_{ij} is the co-factor of the element in the ith row and jth column of the determinent

$$\Delta = \begin{vmatrix} 1 & r_{12} & r_{13} & \ldots & \ldots & r_{1n} \\ r_{21} & 1 & r_{23} & \ldots & \ldots & r_{2n} \\ | & | & | & & & \\ r_{n_1} & r_{n_2} & r_{n_3} & & & 1 \end{vmatrix} \quad \ldots\ldots\ldots(9.31)$$

Properties of the Residuals :

1. The sum of product of corresponding values of a variate and a residual is zero, provided the subscripts of the variate occurs among the secondary subscripts of the residual.

Let the equation of the plane of regression of x_1 on x_2 and x_3 be $x_1 = b_{12.3} x_2 + b_{13.2} x_3$. The normal equations for determining b's give,

$\Sigma x_2 x_{1.23} = 0 = \Sigma x_3 x_{1.23}$.

Similarly from the regression plane of x_2 on x_1 and x_3 and that of x_3 on x_1 and x_2, we have,

$\Sigma x_1 x_{2.13} = 0 = \Sigma x_3 x_{2.13}$ and $\Sigma x_1 x_{3.12} = 0 = \Sigma x_2 x_{3.12}$.

2. The sum of product of two residuals is unaltered by omitting from one residual any or all of the secondary subscripts which are common to both.

Writting $x_{1.2} = x_1 - b_{12}x_2$ we get,

$\Sigma x_{1.23}x_{1.2} = \Sigma x_{1.23}(x_1 - b_{12}x_2) = \Sigma x_{1.23}x_1$.

and $\Sigma x_{1.23}x_{1.23} = \Sigma x_{1.23}(x_1 - b_{12.3}x_2 - b_{13.2}x_3) = \Sigma x_{1.23}x_1$.

3. The sum of product of two residuals is zero provided all the subscripts of residual occur among the secondary subscripts of the second.

By virtue of normal equations, we have

$$\Sigma x_{3.2}x_{1.23} = \Sigma(x_3 - b_{32}x_2)\, x_{1.23} = 0$$

similarly, $\Sigma x_{2.3}x_{1.23} = 0$.

Variance of residuals : Let us consider the plane of regression of x_1 on x_2 and x_3, viz. $x_1 = b_{12.3}x_2 + b_{13.2}x_3$, provided the variables are measured from their means.

The residual is $x_{1.23} = (x_1 - b_{12.3}x_2 - b_{13.2}x_3)$.

Now we shall have to obtain the from of the variance of $x_{1.23}$ which we shall denote by $s^2_{1.23}$ in terms of s_1^2 and correlation co-efficients where s_1^2 is the variance of x_1.

We have $Ns^2_{1.23} = \Sigma x^2_{1.23} = \Sigma x_{1.23}x_1$ vide property 2.

$$= \Sigma x_1(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$

$$= Ns^2_1 - Nb_{12.3}s_1s_2r_{12} - Nb_{13.2}s_1s_3r_{13}.$$

or, $s_1\left(1 - \dfrac{s^2_{1.23}}{s^2_1}\right) = b_{12.3}s_2r_{12} + b_{13.2}s_3r_{13}$.

Eliminating $b_{12.3}$ and $b_{13.2}$ from this equation and normal equation in (9.26) we have,

$$\begin{vmatrix} 1 - \dfrac{s^2_{1.23}}{s^2_1} & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 0$$

or, $\quad \Delta - \dfrac{s^2_{1.23}}{s^2_1}\Delta_{11} = 0$, or, $\dfrac{s^2_{1.23}}{s^2_1} = \dfrac{\Delta}{\Delta_{11}}$

or, $s^2_{1.23} = \dfrac{\Delta}{\Delta_{11}} s_1^2$ ....(9.32)

where Δ and Δ_{11} are defined in (9.29).

Remark : In general, for the distribution of n variates,

$$s^2_{1.23......n} = s^2_1 \frac{\Delta}{\Delta_{11}}$$ where Δ and Δ_{11} are defined in (9.31).

Example 9.13 Find the regression equation of x_1 on x_2 and x_3 from the following results.

Variate	Means	St. deviation	
x_1	28.02	4.42	$r_{12} = 0.80$.
x_2	4.91	1.10	$r_{13} = -0.40$
x_3	594	85	$r_{23} = -0.56$.

Solution : We know the following regression equation of x_1 on x_2 and x_3,

$$(x_1 - 28.02) = -\frac{s_1 \Delta_{12}}{s_2 \Delta_{11}}(x_2 - 4.91) - \frac{s_1 \Delta_{13}}{s_3 \Delta_{11}}(x_3 - 594)$$

where $\Delta_{11} = -\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} = 1 - r^2_{23} = 1 - (0.56)^2 = 0.681.$

where $\Delta_{12} = -\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13}\ r_{23} - r_{21} = -0.57.$

and $\Delta_{13} = \begin{vmatrix} r_{21} & 1 \\ r_{13} & r_{23} \end{vmatrix} = r_{12}r_{23} - r_{13} = -0.048.$

Therefore, the regression plane is given by

$$(x_1 - 28.02) = -\frac{4.42}{1.10} \times \frac{0.57}{0.681}(x_2 - 4.91) - \frac{4.42}{85} \times \frac{-0.048}{0.681} (x_3 - 594)$$

$$= \frac{2.5194}{0.7491} (x_2 - 4.91) + \frac{0.2122}{57.885}(x_3 - 594)$$

$$= 3.36(x_2 - 4.91) + 0.004(x_3 - 594)$$

or, $x_1 - 3.36x_2 - 0.004x_3 - 9.15 = 0$

Partial Correlation Co-efficient : When there are more than two variables, product moment correlation co-efficient between two variables may give partial information. In such situation, one may want to know the correlation co-efficient between two variables x_1 and x_2 when the effects of x_3, x_4 etc. on x_1 and x_2 are eliminated. This correlation is known as partial correlation and the correlation co-efficient between x_1 and x_2 when the linear effect of the other variables on them has been eliminated is called partial correlation co-efficient, and is denoted by $r_{12.34.......}$

Let us consider three variables x_1, x_2 and x_3. $x_{1.3} = x_1 - b_{13}x_3$ may be considered as that part of the variable x_1 after eliminating the effect of x_3, similarly $x_{2.3}$ can be defined also.

Therefore, $r_{12.3} = \dfrac{\Sigma x_{1.3}.x_{2.3}}{\sqrt{(\Sigma x^2_{1.3})(\Sigma x^2_{2.3})}}$

Now $\Sigma x_{1.3}x_{2.3} = \Sigma[(x_1 - b_{13}x_3)(x_2 - b_{23}x_3)]$

$= \Sigma x_1x_2 - b_{23}\Sigma x_1x_3 - b_{13}\Sigma x_2x_3 + b_{13}b_{23}\Sigma x^2_3$

$= Ns_1s_2r_{12} - b_{23}Ns_1s_3r_{13} - b_{13}\,Ns_2s_3r_{23} + b_{13}b_{23}Ns^2_3$

Putting $b_{23} = \dfrac{s_2}{s_3}r_{23}$ and $b_{13} = \dfrac{s_1}{s_3}r_{13}$ we get,

$$\Sigma x_{1.3}x_{2.3} = N(r_{12} - r_{13}r_{23})s_1s_2.$$

Again, $\Sigma x^2_{1.3} = \Sigma x_{1.3}x_{1.3} = \Sigma x_1x_{1.3} = \Sigma x_1(x_1 - b_{13}x_3)$

$= \Sigma x^2_1 - b_{13}\Sigma x_1x_3 = Ns_1^2 - Nb_{13}s_1s_3r_{13} = Ns_1^2(1 - r_{13}^2).$

Similarly $\Sigma x_{2.3}^2 = Ns_2^3(1 - r_{23}^2)$.

$$\text{Now, } r_{12.3} = \frac{Ns_1s_2(r_{12} - r_{13}r_{23})}{\sqrt{Ns_1^2(1-r_{13}^2)Ns_2^2(1-r_2s^2)}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \quad ....(9.33)$$

$r_{1.23}$ can also be obtained in terms of minors of the determined Δ as defined earlier. $b_{12.3}$ is the regression co-efficient of $x_{1.3}$ on $x_{2.3}$ similarly $b_{21.3}$ is the regression co-efficient of $x_{2.3}$ on $x_{1.3}$. Since we know that the correlation co-efficient is the geometric mean of the regression co-efficients. then, .

$r^2_{12.3} = b_{12.3} \times b_{21.3}$.

Putting the value of $b_{12.3}$ and $b_{21.3}$ we have, $r^2_{12.3} = \dfrac{\Delta^2_{12}}{\Delta_{11}\Delta_{22}}$, Since, $\Delta_{12} = \Delta_{21}$.

$$\therefore\ r_{12.3} = \frac{\Delta_{12}}{\sqrt{\Delta_{11}\Delta_{22}}}. \quad ....(9.34)$$

This formula is convenient to get the partial correlation co-efficient of higher order.

Example 9.14 If all the correlation co-efficnet of zero order in a set of P-variates are equal to ρ. Show that every partial correlation co-efficient of sth orders $\dfrac{\rho}{1 + s\rho}$

Solution : We are given that $r_{mn} = \rho$; $\quad m, n = 1,2......P$, $m \neq n$.

We have partial correlation co-efficient of first order.

$$r_{ij.k} = \frac{r_{ij} - r_{.jk}\, r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}} = \frac{\rho - \rho^2}{\sqrt{(1-\rho^2)(1-\rho^2)}} = \frac{\rho}{1+\rho}.$$

Partial correlation co-efficients of second order are given by,

$$r_{ij.kl} = \frac{r_{ij.k} - r_{ik.1}\, r_{jk.1}}{\sqrt{(1 - r^2_{jk.1})(1 - r^2_{jk.1})}} = \frac{\left(\frac{\rho}{1+\rho}\right) - \left(\frac{\rho}{1+\rho}\right)^2}{1 - \left(\frac{\rho}{1+\rho}\right)^2}$$

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

Thus every partial correlation co-efficient of second order is given by,

$$\left(\frac{\rho}{1+2\rho}\right)$$

Proceeding this way i.e. by the method of induction every partial correlation co-efficient of sth order is given by $\frac{\rho}{(1+s\rho)}$. Hence proved.

Example 9.15 From a hypothetical data of three related variables x_1, x_2 and x_3, it is obtained that $r_{12} = 0.59$, $r_{13} = 0.46$ and $r_{23} = 0.77$

where r_{ij} is the correlation co-efficient between x_i and x_j ; $i, j = 1, 2, 3$ $i \neq j$. Find. partial correlation co-efficient $r_{12.3}$.

Solution : Partial correlation co-efficient,

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}} = \frac{0.59 - 0.46 \times 0.77}{\sqrt{(1 - 0.46^2)(1 - 0.77^2)}}$$

$$= \frac{0.536}{\sqrt{(1 - 0.2116)(1 - 0.5929)}} = 0.95 \text{ app.}$$

Example 9.16 Prove the identity, $\quad b_{12.3} \times b_{23.1} \times b_{31.2} = r_{12.3} \times r_{23.1} \times r_{31.2}$.

Solution : We know $b_{12.3} = \frac{s_{1.3}}{s_{2.3}} r_{12.3}$. $b_{23.1} = \frac{s_{2.1}}{s_{3.1}} r_{23.1}$ and $b_{31.2} = \frac{s_{3.2}}{s_{1.2}} r_{31.2}$.

$\therefore$ $b_{12.3} \times b_{23.1} \times b_{31.2} = r_{12.3} \times r_{23.1} \times r_{31.2} \frac{s_{1.3} \times s_{2.1} \times s_{3.2}}{s_{2.3} \times s_{3.1} \times s_{1.2}}$

$= r_{12.3} \times r_{23.1} \times r_{31.2}$. since $s_{1.3} = s_{1.2}$, $s_{2.3}$ $s_{2.3} = s_{2.1}$ and $s_{3.1} = s_{3.2}$.

Hence the result.

Multiple Correlation Co-efficient : Here also we consider tri-variate distribution in which each of the variables x_1, x_2 and x_3 has N observations. $x_{1.23}$ is the multiple regression of x_1 on x_2 and x_3.

Then the correlation co-efficient between x_1 and the expected value of the variate is called the multiple correlation co-efficient, denoted by $R_{1.23}$.

We know the expected value of x_1 as X_1 which is $X_1 = (x_1 - x_{1.23})$.

Therefore, $R_{1.23} = \frac{\Sigma x_1 X_1}{\sqrt{(\Sigma x_1^2)(\Sigma X_1^2)}}$

Now we have, $\Sigma x_1 X_1 = \Sigma x_1(x_1 - x_{1.23}) = \Sigma x_1^2 - \Sigma x_1 x_{1\cdot 23}$

$= \Sigma x^2_1 - \Sigma x^2_{1.23} = Ns_1^2 - Ns^2_{1.23}$.

Also $\quad \Sigma X^2_1 = \Sigma(x_1 - x_{1.23})^2$

$= \Sigma x^2_1 - 2\Sigma x_1 x_{1.23} + \Sigma x^2_{1.23}$. Since, $\Sigma x_1 x_{1.23} = \Sigma x^2_{1.23}$

$= \Sigma x^2_1 - \Sigma x^2_{1.23}$, $Ns_1{}^2 - Ns^2_{1.23}$.

Therefore, $R_{1.23} = \frac{s^2_1 - s^2_{1.23}}{s_1 \sqrt{s^2_1 - s^2_{1.23}}} = \frac{\sqrt{s^2_1 - s^2_{1\cdot 23}}}{s_1}$

$= \sqrt{\frac{s^2_1 - s^2_{1.23}}{s^2_1}} = \left(1 - \frac{s^2_{1.23}}{s^2_1}\right)^{\frac{1}{2}}$.

or, $R^2_{1.23} = 1 - \frac{s^2_{1.23}}{s^2_1} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$

or, $1 - R^2_{1.23} = \frac{\Delta}{\Delta_{11}}$ ....(9.35)

where Δ and Δ_{11} are defined in (9.29). This formula is used for calculating multiple correlation co-efficient for more than three variates.

Example 9.17 Multiple correlation co-efficient can be expressed in terms of total and partial correlation co-efficient i. e. $1 - R^2_{1.23} = (1 - r^2_{12})(1 - r^2_{13.2})$.

Solution : We have, $R^2_{1.23} = 1 - \frac{\Delta}{\Delta_{11}}$

$$\text{or, } 1 - R^2_{1.23} = \frac{\Delta}{\Delta_{11}} = \frac{1 - r^2_{12} - r^2_{23} - r^2_{13} + 2r_{12}r_{23}r_{31}}{1 - r^2_{23}} \qquad .....(9.36)$$

$$\text{Also we know, } r^2_{13.2} = \frac{\Delta^2_{13}}{\Delta_{11}\Delta_{33}} = \frac{(r_{13} - r_{12}r_{32})^2}{(1 - r^2_{12})(1 - r^2_{32})}$$

$$\text{or, } 1 - r^2_{13.2} = 1 - \frac{(r_{13} - r_{12}r_{32})^2}{(1 - r^2_{12})(1 - r^2_{23})}$$

$$= \frac{1 - r^2_{12} - r^2_{13} - r^2_{23} + 2r_{12}r_{23}r_{31}}{(1 - r^2_{12})(1 - r^2_{32})}$$

$$\text{or, } (1 - r^2_{12})(1 - r^2_{13.2}) = \frac{1 - r^2_{12} - r^2_{13} - r^2_{32} + 2r_{12}r_{23}r_{31}}{(1 - r^2_{32})} \qquad .....(9.37)$$

Companing R. H. S of (9.36) and (9.37) we have,

$(1 - R^2_{1.23}) = (1 - r^2_{12})(1 - r^2_{13.2})$. Hence the result.

Remarks :

1) In general, for a n variates we have, $1 - R^2_{1.23....n} = \frac{\Delta}{\Delta_{11}}$

where Δ and Δ_{11} are defined in (9.31).

2) $R_{1.23}$ is simple correlation between x_1 and its expected value X_1, hence its should lie between - 1 and + 1. But since $\Sigma x_i X_1 = \Sigma x^2_1$ which cannot be negative. Hence $R_{1.23}$ is necessarily positive or zero, that is why we conclude, $0 \leq R_{1.23} \leq 1$.

3) If $R_{1.23} = 1$, the association is perfect and all the residuals are zero, and as such $s^2_{1.23} = 0$, the observed and the expected values of x_1 coincides. Therefore, we can conclude that x_1 is perfectly linear function of x_2 and x_3.

4) If $R_{1.23} = 0$, X_1 is completely uncorrelated with x_1 and thus the multiple regression equation fails to throw any light on the value of x_1 when x_2 and x_3 are known.

Example 9.18 Calculate multiple correlation co-efficient from the data given in Example 9.15.

Solution : We know, $R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$

$$= \frac{(0.59)^2 + (0.46)^2 - 2 \times 0.59 \times 0.46 \times 0.77}{1 - (0.77)^2} = \frac{0.1417}{0.4071} = 0.3456$$

$\therefore R_{1.23} = 0.584$ (app).

Example 9.19 If all the correlation co-efficient of zero order in a set of p variates are equal to ρ, show that the multiple correlation co-efficient of a variate with other (p - 1) variate is given by,

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

Solution : We know that, $1 - R^2 = \dfrac{\Delta}{\Delta_{11}}$

where $\Delta = \begin{vmatrix} 1 & \rho & \rho \ldots\ldots\rho \\ \rho & 1 & \rho \ldots\ldots\rho \\ \vdots & \vdots & \vdots \qquad \vdots \\ \rho & \rho & \rho \ldots\ldots 1 \end{vmatrix}$ a determinent of order p.

and $\Delta_{11} = \begin{vmatrix} 1 & \rho \ldots\ldots\rho \\ \rho & 1 \ldots\ldots\rho \\ \vdots & \vdots \qquad \vdots \\ \rho & \rho \ldots\ldots 1 \end{vmatrix}$ a determinent of order (p-1).

We have,

$\Delta = \{1 + (p-1)\rho\} \begin{vmatrix} 1 & \rho & \rho \ldots\ldots\rho \\ 1 & 1 & \rho \ldots\ldots\rho \\ \vdots & \vdots & \vdots \qquad \vdots \\ 1 & \rho & \rho \ldots\ldots 1 \end{vmatrix}$ adding $c_2, c_3 \ldots\ldots c_p$ to c_1 where c_i indicates ith column.

$= \{1 + (p-1)\rho\} \begin{vmatrix} 1 & \rho & \rho \ldots\ldots\ldots\ldots & \rho \\ 0 & (1-\rho) & \ldots\ldots\ldots & 0 \\ 0 & 0 & (1-\rho) & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 \ldots & (1-\rho) \end{vmatrix}$ on operating $R_i - R_1$, $i = 2\ 3 \ldots\ldots p$ where R_i indicates ith row.

$\therefore \Delta = \{1 + (p-1)\rho\}(1-\rho)^{p-1}$

Similarly we can get, $\Delta_{11} = \{1 + (p-2)\rho\}(1-\rho)^{p-2}$.

Therefore, $1 - R^2 = \dfrac{\Delta}{\Delta_{11}} = (1-\rho)\left[\dfrac{1 + (p-1)\rho}{1 + (p-2)\rho}\right]$. Hence shown.

10 EXACT SAMPLING DISTRIBUTION OF χ^2 (CHI-SQUARE), STUDENT'S t, F AND TEST OF SIGNIFICANCE.

10.1 Random Sampling

In Chapter 2, we have discussed the terms—population and sample. It is obvious from the discussion that sample is necessary to ascertain the properties and characteristics of the population. For this purpose random samples are essential. A random sample is one in which each unit of population has an equal chance of being included in it and the procedure to have such a sample is known as random sampling.

Parameter, Statistic and its Sampling Distribution : For drawing valid inference about the population we, in practice, deal with samples and obtain the estimates of the population characteristics. The unknown characteristics of the population are usually called parameters. And the estimate of a certain parameter is called statistic. A statistic is generally a function of a set of sample values. It may be pointed out that there may be a number of choices of the samples that can be drawn from the population. Hence the statistic itself is liable to vary from one sample to another. These differences in the values of the statistic are called sampling fluctuation. If the number of samples each of size n say, are taken from the same population and for each sample the value of the statistic can be calculated, a series of values of the statistic can be obtained. For large number of samples each of size n, a frequency table can be constructed from the series of statistic, giving us the sampling distribution of the statistic. In case of random sampling, the sampling distribution of the statistic can be obtained in probabilistic sense if the nature of the parent population is given or known. Thus the sampling distribution is defined as the probability distribution of a statistic derived from random samples drawn from some specified parent population.

Like any other distributions, a sampling distribution may have mean standard deviation and moments of higher order. The standard deviation of the statistic is usually called standard error of the statistic.

We shall derive the sampling distribution of the χ^2 statistic, t-statistic and F-statistic and also indicate their properties, uses and applications in the next sections.

Degrees of Freedom (d. f) : The number of degrees of freedom (d. f.) is equal to the number of independent comparisons between the observations of a sample. If there is a sample of size n, (n -1) independent comperisons can be made and therefore the d. f. is (n - 1). Again if the sample is arranged in k classes, then the d. f. is (k -1) as (k -1) frequencies are specified, the other is determined by the total size n. Thus if b functions of the sample values are held constant the number of d. f. is reduced by b.

10.2 χ^2-(Chi-square) Distribution

χ^2-distribution with n d. f. is the distribution of the sum of squares of n independent standardised normal variates.

Let $x_1, x_2, \ldots\ldots x_n$ be n independent N(0,1) variates then the statistic χ^2 is defined by $\chi^2 = \sum_{i=1}^{n} x_i^2$ and the distribution of $\sum_{1=1}^{n} x_i^2$ is χ^2- distribution with n d. f. usually denoted by χ^2_n. If x_i's are independent $N(\mu_i, \sigma 2_i)$ variates then

$$\sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 \text{ is also a } \chi^2_n.$$

Derivation of χ^2-distribution : Let $x_1, x_2, \ldots\ldots x_n$ be n independent $N(\mu_i, \sigma^2_i)$ variates, we are to find the distribution of

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n} u^2_i \text{ where } u_i = \frac{x_i - \mu_i}{\sigma_i}$$

Since x_i's are independent, u_i's are also independent.

$$\text{Therefore, } \varphi_{\chi^2}(t) = \varphi_{\sum_{i=1} u^2_i}(t) = \prod_{i=1}^{n} \varphi_{u^2_i}(t) = [\varphi_{u^2_i}(t)]^n$$

where φ indicates the characteristic function.

Now, since we know that, $\varphi_{u^2_i}(t) = E[e^{itu_i^2}]$

$$= \int_{-\infty}^{\infty} e^{itu_i^2} f(u_i)\,du_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itu_i^2} e^{-\frac{1}{2}u_i^2}\,du_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itu_i^2(1-2it)}\,du_i = \frac{1}{(1-2it)^{\frac{1}{2}}}$$

Since $\dfrac{1}{\sqrt{2\pi}\,(1-2it)^{\frac{1}{2}}} \displaystyle\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2(1-2it)}\,du = 1$.

Now, $\varphi_{\chi^2}(t) = \dfrac{1}{(1-2it)^{\frac{n}{2}}}$ which is the characteristic function of Gamma variate with parameters $\left(\frac{1}{2}, \frac{n}{2}\right)$. Hence from the uniqueness theorem of characteristic function the p.d. f. of χ^2_n is

$$f(\chi^2_n) = \frac{1}{2^{\frac{n}{2}}\Gamma\frac{n}{2}} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}; \quad 0 \le \chi^2 \le \infty. \qquad .....(10.1),\ )$$

Remarks :

Normal distributution is a particular case of χ^2-distribution with n = 1.

If x_i (i = 1, 2,......n) be n independent normal variate with sample mean $\overline{x}$ and known population variance σ^2 then $\dfrac{(n-1)\,S^2}{\sigma^2}$ is a χ^2 variate with (n - 1) d. f. where $S^2 = \dfrac{1}{n-1}\Sigma(x_i - \overline{x})^2$.

iii) $x_i \sim N(\mu, \sigma^2)$ variate then $\overline{x} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$

hence $\left[\dfrac{\overline{x}-\mu}{\sigma/\sqrt{n}}\right]^2$ is a χ^2 variate with 1. d. f.

χ^2 - Probability Curve : For different values of n, the degrees of freedom the different types of curves can be obtained as shown in Fig. 10.1.

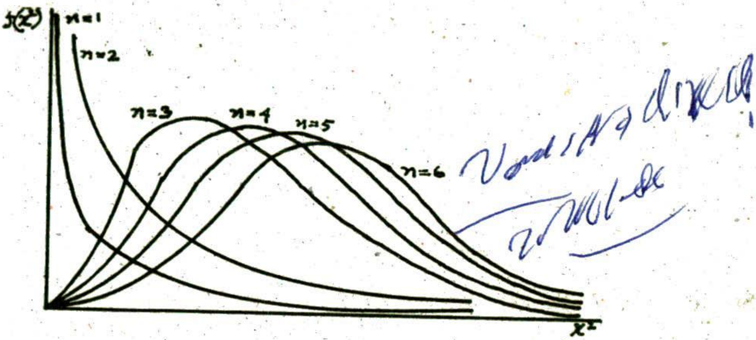


Fig. 10.1 χ^2-Probability curves.

Moment generating function of χ^2-distribution :

$$M\chi^2(t) = E[e^{t\chi^2}] = \int_0^\infty e^{t\chi^2} f(\chi^2)\, d\chi^2.$$

$$= \frac{1}{2^{\frac{n}{2}} \Gamma\frac{n}{2}} \int_0^\infty e^{t\chi^2} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1} d\chi^2.$$

$$= \frac{1}{2^{\frac{n}{2}} \Gamma\frac{n}{2}} \int_0^\infty e^{-\frac{\chi^2}{2}(1-2t)} (\chi^2)^{\frac{n}{2}-1} d\chi^2.$$

$$= \frac{1}{(1-2t)^{\frac{n}{2}}} \int_0^\infty \frac{1}{2^{\frac{n}{2}} \Gamma\frac{n}{2}} \int_0^\infty e^{-\frac{\chi'^2}{2}} (\chi'^2)^{\frac{n}{2}-1} d\chi'^2. \text{ where } \chi'^2 = \chi^2(1-2t).$$

$$= \frac{1}{(1-2t)^{\frac{n}{2}}} \qquad \text{.........(10.2)}$$

First four moments of χ^2 distribution :

We know, $M\chi^2_n(t) = (1-2t)^{-\frac{n}{2}}$

Expanding we get, $M\chi^2_n(t) = 1 + \frac{n}{2}(2t) + \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)}{2!}(2t)^2 + \ldots\ldots$

$$\ldots\ldots + \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)\cdot\left(\frac{n}{2}+2\right)\ldots\ldots\left(\frac{n}{2}+r-1\right)}{r!}(2t)^r + \ldots\ldots \qquad \text{..(10.3)}$$

As μ'_r = Co-efficient of $\frac{t^r}{r!}$ in the expansion of $M\chi^2_n(t)$ in (10.3).

Therefore, $\mu_1' = n,$

$$\mu_2' = n(n+2)$$

Since, $\mu_2 = \mu_2' - \mu_1'^2 = 2n.$

Again, $\mu_3' = n(n+2)(n+4)$

Since, $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 8n.$

And again, $\mu_4' = n(n+2)(n+4)(n+6).$

Since, $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 48n + 12n^2.$

We know, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{8}{n}$ and $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{12}{n} + 3.$

Remarks :

As $n \to \infty$, $\beta_1 \to 0$ and $\beta_2 \to 3$, hence χ^2 - distribution tends to normal distribution if the degrees of freedom, n is very large.

Additive property of χ^2 - variate : If $\chi^2 n_1$ and $\chi^2 n_2$ are two independent χ^2 - variates with n_1 and n_2 d. f. respectively then $\chi^2 n_1 + \chi^2 n_2$ is also a χ^2 - variate with $(n_1 + n_2)$ d. f.

Proof : We know, $M\chi^2 n_1(t) = (1-2t)^{-\frac{n_1}{2}}$ and $M\chi^2 n_2(t) = (1-2t)^{-\frac{n_2}{2}}.$

Since $\chi^2 n_1$ and $\chi^2 n_2$ are independent then,

$$M\chi^2 n_1 + \chi^2 n_2(t) = (1-2t)^{-\frac{(n_1+n_2)}{2}} \qquad \text{.........(10.4)}$$

which is the moment generating function of a χ^2 - variate with $(n_1 + n_2)$ d.f. Hence proved.

The result can be extended for any number of independent χ^2 - variates.

Remarks : The converse of the result is also true i. e. if χ^2_i (i = 1, 2,.....k) are χ^2 - variate with n_i (i = 1, 2,........k) d. f. respectively then $\sum_{i=1}^{k} \chi^2_i$ is a χ^2 variate with $\sum n_i$ d. f. then χ^2_i's are independent.

Another useful version of the converse is as follows : If X and Y are two non-negative variates such that X + Y follows χ^2 - distribution with $(n_1 + n_2)$ d. f. and if any one of them, say X is χ^2 - variate with n_1 d. f. then the rest one Y is also a χ^2 with n_2 d.f. The above version is true for any number of such variates.

Theorem 10.1 If $\chi^2 n_1$ and $\chi^2 n_2$ are two independent χ^2 - variates with n_1 and n_2 d. f. respectively then $\frac{\chi^2 n_1}{\chi^2 n_2}$ is a $\beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate.

Proof : Since $\chi^2 n_1$ and $\chi^2 n_2$ are independent χ^2 - variate with n_1 and n_2 d. f. respectively then the joint probability differential is given by the multiplicative law of probability as shown below :

$$dF(\chi^2 n_1 \chi^2 n_2) = dF(\chi^2 n_1)\, dF(\chi^2 n_2).$$

$$= \frac{1}{2^{\frac{n_1+n_2}{2}} \Gamma\frac{n_1}{2} \Gamma\frac{n_2}{2}} e^{-\frac{\chi^2 n_1 + \chi^2 n_2}{2}} (\chi^2 n_1)^{\frac{n_1}{2}-1} (\chi^2 n_2)^{\frac{n_2}{2}-1} d\chi^2 n_1 d\chi^2 n_2,$$

$0 \le (\chi^2 n_1, \chi^2 n_2) \le \infty.$

Let us put, $u = \frac{\chi^2 n_1}{\chi^2 n_2}$ and $v = \chi^2{}_{n_2}$.

So that $uv = \chi^2 n_1$ and $\chi^2 n_2 = v$.

Jacobian of transformation is given by

$$|J| = \begin{vmatrix} \frac{d\chi^2 n_1}{du} & \frac{d\chi^2 n_1}{dv} \\ \frac{d\chi^2 n_2}{du} & \frac{d\chi^2 n_2}{dv} \end{vmatrix} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v.$$

$\therefore d\chi^2 n_1 d\chi^2 n_2 = v\, du\, dv.$

Then the joint distribution of u and v becomes

$$d\,G(u,v) = \frac{1}{2^{\frac{n_1+n_2}{2}} \Gamma\frac{n_1}{2} \Gamma\frac{n_2}{2}} e^{-\frac{v(1+u)}{2}} (uv)^{\frac{n_1}{2}-1} v^{\frac{n_2}{2}-1} v\, du\, dv.$$

$$= \frac{1}{2^{\frac{n_1+n_2}{2}}\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} e^{-\frac{v(1+u)}{2}} u^{\frac{n_1}{2}-1} v^{\frac{n_1+n_2}{2}-1} du\,dv;\quad 0 \le (u,v) \le \infty.$$

We know that integrating $dG(u,v)$ with respect to v over range 0 to ∞, we get

$$dG(u) = \int_0^{\infty} dG(u,v)\,dv.$$

$$= \frac{1}{2^{\frac{n_1+n_2}{2}}\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} u^{\frac{n_1}{2}-1} du \int_0^{\infty} e^{-\frac{v(1+u)}{2}} (v)^{\frac{n_1+n_2}{2}-1} dv.$$

$$= \frac{u^{\frac{n_1}{2}-1}}{2^{\frac{n_1+n_2}{2}}\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} \frac{\Gamma\frac{n_1+n_2}{2}}{\left(\frac{1+u}{2}\right)^{\frac{n_1+n_2}{2}}} du.$$

$$= \frac{1}{\beta\left(\frac{n_1}{2},\frac{n_2}{2}\right)} \frac{u^{\frac{n_1}{2}-1}}{(1+u)^{\frac{n_1+n_2}{2}}} du \qquad ;0 \le u \le \infty. \qquad \text{........(10.5)}$$

Hence $u = \frac{\chi^2 n_1}{\chi^2 n_2}$ is a $\beta_2\left(\frac{n_1}{2},\frac{n_2}{2}\right)$ variate.

Theorem 10.2 If $\chi^2 n_1$ and $\chi^2 n_2$ are independent χ^2-variates with n_1 and n_2 d. f. respectively, then $u = \frac{\chi^2 n_1}{\chi^2 n_1 + \chi^2 n_2}$ is independently distributed as

$\beta_1\left(\frac{n_1}{2},\frac{n_2}{2}\right)$.

Proof : As given in theorem 10.1 we have the joint probability differential.

$$dP(\chi^2 n_1 \chi^2 n_2) = \frac{1}{2^{\frac{n_1+n_2}{2}}\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} e^{-\frac{\chi^2 n_1 + \chi^2 n_2}{2}} (\chi^2 n_1)^{\frac{n_1}{2}-1}$$

$$(\chi^2 n_2)^{\frac{n_2}{2}-1} d\chi^2 n_1 d\chi^2 n_2; \qquad 0 \le \chi^2 n_i,\ \chi^2 n_2 \le \infty$$

Let us put, $u = \frac{\chi^2 n_1}{\chi^2 n_1 + \chi^2 n_2}$ and $v = \chi^2 n_1 + \chi^2 n_2$

So that $uv = \chi^2 n_1$ and $\chi^2 n_2 = v - \chi^2 n_1 = (1 - u)v$.

Since $\chi^2 n_1$ and $\chi^2 n_2$ range from 0 to ∞, u ranges from 0 to 1 and v ranges from 0 to ∞.

Jacobian transformation, J is given by

$$|J| = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v.$$

Threfore, $d\chi^2 n_1 d\chi^2 n_2 = v\,du\,dv$ and

$$dG(u,v) = \frac{1}{2^{\frac{n_1+n_2}{2}}\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} e^{-\frac{v}{2}} (uv)^{\frac{n_1}{2}-1} \{(1-u)v\}^{\frac{n_2}{2}-1} v\,du\,dv.$$

$$= \frac{1}{2^{\frac{n_1+n_2}{2}}\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} u^{\frac{n_1}{2}-1} (1-u)^{\frac{n_2}{2}-1} e^{-\frac{v}{2}} v^{\frac{n_1+n_2}{2}-1} du\,dv.$$

$$= \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} u^{\frac{n_1}{2}-1} (1-u)^{\frac{n_2}{2}-1} du \frac{1}{2^{\frac{n_1+n_2}{2}}\Gamma\left(\frac{n_1+n_2}{2}\right)}$$

$$\times e^{-\frac{v}{2}} v^{\frac{n_1+n_2}{2}-1} dv.$$

Since the joint probability differential of u and v is the product of their respective probability differential, u and v are independently distributed with,

$$dG_1(u) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} u^{\frac{n_1}{2}-1} (1-u)^{\frac{n_2}{2}-1} du, 0 \le u \le 1. \qquad \text{.........(10.6)}$$

and
$$dG_2(v) = \frac{1}{2^{\frac{n_1+n_2}{2}}\Gamma\frac{n_1+n_2}{2}} e^{-\frac{v}{2}} v^{\frac{n_1+n_2}{2}-1} dv, 0 \le v \le \infty, \qquad \text{....(10.7)}$$

That is, u is a $\beta_1\left(\frac{n_1}{2},\frac{n_2}{2}\right)$ variate and v is a χ^2 - variate with (n_1+n_2) d.f.

Theorem 10.3 For large n, the d.f., show that $\sqrt{2\chi^2 n} \sim N(\sqrt{2n},1)$.

Proof. We know that $E(\chi^2{}_n) = n$. and $V(\chi^2 n) = 2n$. Now let us define

$z = \frac{\chi^2{}_n - n}{\sqrt{2n}}$ which tends to N(0,1) for large n.

Let us consider

$$P\left[\frac{\chi^2 n - n}{\sqrt{2n}} \le z\right] = P[\chi^2 n \le n + z\sqrt{2n}]$$

$$= P[2\chi^2 n \le 2n + 2z\sqrt{2n}]$$

$$= P[\sqrt{2\chi^2 n} \le (2n + 2z\sqrt{2n})^{\frac{1}{2}}]$$

$$= P[\sqrt{2\chi^2 n} \le \sqrt{2n}(1 + z\sqrt{\tfrac{2}{n}})^{\frac{1}{2}}]$$

$$= P\left[\sqrt{2\chi^2 n} \le \sqrt{2n}\left(1 + \frac{z}{\sqrt{2n}} - \frac{z^2}{4n} + \ldots\right)\right]$$

$$= P[\sqrt{2\chi^2 n} \le \sqrt{2n} + z] \text{ for large n}$$

$$P[\sqrt{2\chi^2 n} - \sqrt{2n}] \le z] \text{ for large n.}$$

As we know, for large n, $\frac{\chi^2{}_n - n}{\sqrt{2n}} \sim N(0,1)$.

we conclude that $\sqrt{2\chi^2 n} - \sqrt{2n} \sim N(0,1)$ for large n,

which implies that $\sqrt{2\chi^2 n}$ is asymptotically $N(\sqrt{2n}, 1)$.

Remark : The above approximation is valid for $n \ge 30$. For moderate n, R. A. Fisher has proved that the approximation is improved by taking $\sqrt{(2n-1)}$ instead of $\sqrt{2n}$.

Theorem 10.4 If the variable $x_1, x_2, \ldots\ldots x_n$ are independently distributed in the rectangular form $dF = dx$, $0 \le x \le 1$, then

$-2\log(x_1 x_2 \ldots\ldots x_n)$ is distributed as χ^2 with 2n d.f.

Proof : Let $-2\log(x_1, x_2 \ldots x_n) = p_1 + p_2 + \ldots + p_n$

where $p_i = -2\log x_i$, $i = 1, 2 \ldots n$

or, $x_i = e^{-\frac{p_i}{2}}$

The probability function of p_i is given by, $f(p_i) = f(x_i) \left| \frac{dx_i}{dp_i} \right|$

Since $dF(x) = dx$, $f(x) = 1$ for all x within the range 0 to 1.

$$\therefore f(p_i) = 1 \left| \frac{1}{2} e^{-\frac{p_i}{2}} \right| = \frac{1}{2} e^{-\frac{p_i}{2}}$$

which is the probability function of χ^2 - distribution with 2 d. f.

Therefore, by the additive property of χ^2 - distribution,

$$2\log(x_1 x_2 \ldots x_n) = \sum_{i=1}^{n} p_i \text{ distributed as } \chi^2 \text{ with } 2n \text{ d. f.}$$

10.3 Student's t-Distribution

W. S. Gosset (1908) under the penname of Student defined the t-statistic with (n -1) d. f. by

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (10.8)$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ and μ is the population mean

He derived the distribution of t which is known as Student's t-distribution.

Fisher (1926) defined t-statistic with δ d. f. as the ratio of a standardised normal variate to the square root of an independent Chi-square variate divided by its d.f. δ. That is, he defined $t = \frac{u}{\sqrt{\frac{\chi^2_\delta}{\delta}}}$ with δ d. f. where u is a N(0,1) variate, and χ^2 is a chi-square variate with δ d. f.

Theorem 10.5 The value of Fisher's t is same as Student's t.

Proof : Let $x_1, x_2, \ldots x_n$ be n independent $N(\mu, \sigma^2)$ variates, then

$u = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is a N (0,1) variate, $\dfrac{(n-1)S^2}{\sigma^2}$ is distributed as χ^2 with (n - 1) d. f.

where $S^2 = \dfrac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$, then the statistic, $t = \dfrac{u}{\sqrt{\dfrac{\chi^2_\delta}{\delta}}}$

$$= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = \frac{\bar{x} - \mu}{S/\sqrt{n}} \qquad \text{.........(10.9)}$$

which is same as in (10.8).

The d. f. of Fisher's t is same as the d. f. of chi-square variate and this is more general than the Student's-t.

Derivation of Fisher's t-distribution : From (10.8) we have

$$t^2 = \frac{n(\bar{x} - \mu)^2}{S^2} = \frac{n(\bar{x} - \mu)^2}{\frac{ns^2}{(n-1)}}; \text{ Since } ns^2 = (n-1)S^2 = \sum_{i=1}^{n} (xi - \bar{x})^2$$

or, $\dfrac{t^2}{(n-1)} = \dfrac{(\bar{x} - \mu)^2}{s}$ or , $\dfrac{t^2}{\delta} = \dfrac{(\bar{x} - \mu)^2}{\sigma^2/n} \Big/ \dfrac{ns^2}{\sigma^2}$ where $\delta = (n-1)$.

Since $x_1, x_2, \ldots\ldots x_n$ be a random sample from a normal population with mean μ and variance σ^2, then $\bar{x} \sim N(\mu, \sigma^2/n)$ and $\dfrac{(\bar{x} - \mu)^2}{\sigma^2/n}$ is a χ^2 with 1 d. f. and also $\dfrac{ns^2}{\sigma^2}$ is a χ^2 with (n - 1) d. f. Further since $\bar{x}$ and s^2 are independently distributed then χ^2_1 and $\chi^2_{(n-1)}$ are also independently distributed.

Threfore, $v = \dfrac{t^2}{\delta}$ is the ratio of two independent χ^2 variates with 1 and $\delta = (n-1)$ d. f. respectively. The ratio gives $\beta_2\left(\frac{1}{2}, \frac{\delta}{2}\right)$ variate given in (10.5) and its distribution is given by

$$dF(v) = \frac{1}{\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} v^{-\frac{1}{2}} (1+v)^{-\frac{\delta+1}{2}} dv, \quad 0 \le v \le \infty.$$

Therefore, $dF(t^2) = \dfrac{1}{\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} \left(\dfrac{t^2}{\delta}\right)^{-\frac{1}{2}} \left(1 + \dfrac{t^2}{\delta}\right)^{-\frac{\delta+1}{2}} \dfrac{dt^2}{\delta}$

$$= \frac{1}{\sqrt{\delta}\,\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} (t^2)^{-\frac{1}{2}} \left(1 + \frac{t^2}{\delta}\right)^{-\frac{\delta+1}{2}} dt^2, \quad 0 \le t^2 \le \infty.$$

Now, $dF(t) = \dfrac{1}{\sqrt{\delta}\,\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} t^{-1} \left(1 + \dfrac{t^2}{\delta}\right)^{-\frac{\delta+1}{2}} 2t\,dt.$

$$= \frac{1}{\sqrt{\delta}\,\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} \left(1 + \frac{t^2}{\delta}\right)^{-\frac{\delta+1}{2}} dt, \quad -\infty \le t \le \infty.$$

Thus, $f(t) = \dfrac{\Gamma\left(\frac{\delta+1}{2}\right)}{\sqrt{\pi\delta}\,\Gamma\frac{\delta}{2}} \left(1 + \dfrac{t^2}{\delta}\right)^{-\frac{\delta+1}{2}} dt,$ ....(10.10)

(10.10) is the required p. d. f. of t with $\delta = (n - 1)$ d. f.

t - Probability Curve : The p. d. f. of t-distribution with δ d. f. is

$$f(t) = \frac{1}{\sqrt{\delta}\,\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{\delta}\right)^{\frac{d+1}{2}}}; \quad -\infty \le t \le \infty.$$

it is seen that the curve is symmetrical about the line $t = 0$. Since $f(t) = f(-t)$ As t increase f(t) decreases rapidly and tends to zero at $t \to \infty$.

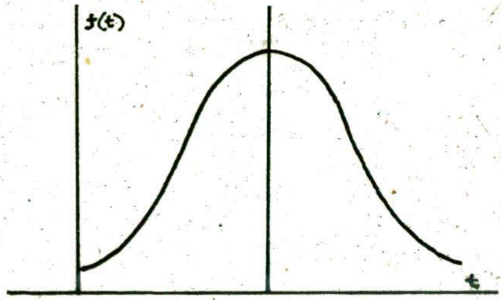


Fig. 10.2 t-Probability curve.

Properties of t-distribution :

Moments : Since f(t) is symmetrical about $t = 0$ all odd order moments about origin vanish, i. e. $\mu'_{2r+1} = 0$; $r = 0, 1, 2......$

In particular, $\mu_1' = 0$ = mean. Hence the central moments coincides with moments about origin i. e. $\mu_{2r+1} = 0$; $r = 0, 1, 2........$

The moments of even order are given by

$$\mu_{2r} = \mu'_{2r} = \int_{-\infty}^{\infty} t^{2r} f(t)\, dt.$$

$$= \frac{2}{\sqrt{\delta}\,\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} \int_0^{\infty} \frac{t^{2r}}{\left(1 + \frac{t^2}{\delta}\right)^{\frac{\delta+1}{2}}} dt.$$

$$= \frac{2}{\sqrt{\delta}\,\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} \int_0^{\infty} \frac{t^{2r}}{\left(1 + \frac{t^2}{\delta}\right)^{\frac{\delta+1}{2}}} \frac{dt^2}{2t}.$$

Let us put $\frac{t^2}{\delta} = y \therefore t^2 = \delta y$

or, $\quad dt^2 = \delta\, dy.$

Since $\quad 0 \le t^2 \le \infty, \qquad 0 \le y \le \infty.$

$$\therefore \mu_{2r} = \frac{2}{\sqrt{\delta}\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} \int_0^{\infty} \delta^r y^r \frac{1}{(1+y)^{\frac{\delta+1}{2}}} \frac{\delta dy}{\sqrt{\delta y}}$$

$$= \frac{\delta^r}{\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} \int_0^{\infty} y^{r-\frac{1}{2}} \frac{1}{(1+y)^{(\frac{\delta}{2}-r)+(r+\frac{1}{2})}} dy$$

$$= \frac{\delta^r}{\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} \beta\left(r + \frac{1}{2}, \frac{\delta}{2} - r\right)$$

$$= \frac{\delta^r (2r-1)(2r-3)\ldots\ldots 3, 1}{(\delta-2)(\delta-4)\ldots\ldots(\delta-2r)}, \frac{\delta}{2} > r \qquad \ldots\ldots\ldots(10.11)$$

Since we know, $\beta(m,n) = \frac{\Gamma m \, \Gamma n}{\Gamma m+n}$

In particular, $\mu_2 = \frac{\delta}{\delta - 2}$ and $\mu_4 = \frac{\delta^2 \times 1 \times 3}{(\delta-2)(\delta-4)} = \frac{3\delta^2}{(\delta-2)(\delta-4)}$

Hence, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$ and $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3(\delta-2)}{(\delta-4)} = 3.\frac{\left(1-\frac{2}{\delta}\right)}{\left(1-\frac{4}{\delta}\right)}$

$=3$; As $\delta \to \infty$ $\beta_2 \to 3$.

Hence for large n, t-distribution tends to normal distribution, as $\delta = (n-1)$

Theorem 10.6 t-distribution tends to standardised normal distribution when the d. f. of t-distribution is large.

Proof : We know,

$$f(t) = \frac{1}{\sqrt{\delta}\,\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)}\left(1+\frac{t^2}{\delta}\right)^{-\left(\frac{\delta+1}{2}\right)}.$$

The constant term, $\frac{1}{\sqrt{\delta}\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)}$

$$= \frac{1}{\sqrt{\delta}} \frac{\Gamma\left(\frac{\delta+1}{2}\right)}{\Gamma\frac{\delta}{2}\Gamma\frac{1}{2}} = \frac{1}{\sqrt{\delta\pi}} \frac{\left(\frac{\delta-1}{2}\right)!}{\left(\frac{\delta}{2}-1\right)!}$$

Using Starling's approximation and taking limit $\delta \to \infty$ we have

$$\frac{1}{\sqrt{\delta}\,\beta\left(\frac{1}{2}, \frac{\delta}{2}\right)} = \frac{1}{\sqrt{\delta\pi}}\left(\frac{\delta}{2}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2\pi}}.$$

Now, $$\underset{\delta \to \infty}{\mathrm{Lt}}\ f(t) = \frac{1}{\sqrt{2\pi}} \underset{\delta \to \infty}{\mathrm{Lt}} \frac{\left(1 + \frac{t^2}{\delta}\right)^{-\frac{\delta}{2}}}{\left(1 + \frac{t^2}{\delta}\right)^{\frac{1}{2}}}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}; \quad -\infty \leq t \leq \infty \qquad \text{.....(10.12)}$$

(10.12) is identical to (8.48)

Since $$\underset{\delta \to \infty}{\mathrm{Lt}} \left(1 + \frac{t^2}{\delta}\right)^{-\frac{\delta}{2}} \to e^{-\frac{t^2}{2}}$$

and $$\underset{\delta \to \infty}{\mathrm{Lt}} \left(1 + \frac{t^2}{\delta}\right)^{\frac{1}{2}} \to 1.$$

Hence it is shown that the distribution of t is N(0, 1) variate for large δ.

10.4 F-Distribution

F-distribution with n_1 and n_2 d. f. is the distribution of the ratio of two independent χ^2s divided by their respective n_1 and n_2 d. f. Thus the F-statistic may be defined as

$F = \dfrac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$ where $\chi^2_{n_1}$ and $\chi^2_{n_2}$ are two independent χ^2 with n_1 and n_2 d. f. respectively. The F-distribution is usually called Snedecor's F-distribution.

Derivation of F-Distribution : Let $\chi^2_{n_1}$ and $\chi^2_{n_2}$ are two independent chi-squares with n_1 and n_2 d. f. respectively then,

$F = \dfrac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$ $\quad \therefore \dfrac{n_1}{n_2} F = \dfrac{\chi^2_{n_1}}{\chi^2_{n_2}}$ being the ratio of two independent chi-square variates with n_1 and n_2 d. f. respectively and is distributed as

$\beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ given in (10.5). Hence the probability function of F is given by,

$$dF(F) = \frac{1}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \frac{\left(\frac{n_1}{n_2}F\right)^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1+n_2}{2}}} d\left(\frac{n_1}{n_2}F\right)$$

$$= \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \frac{F^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1+n_2}{2}}} dF, \quad 0 \le F \le \infty \qquad \text{......(10.13)}$$

Remarks :

$$F = \frac{\chi^2 n_1/n_1}{\chi^2 n_2/n_2} = \frac{\frac{n_1 s_1^2}{\sigma^2}\Big/ n_1}{\frac{n_2 s^2}{\sigma^2}\Big/ n_2} = \frac{s_1^2}{s_2^2}.$$

Thus the distribution of F may be called the distribution of the variance ratio given by Snedecor.

Moments of F-distribution : The rth raw moment is given by

$$\mu'_r = \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \int_0^\infty \frac{F^{\frac{n_1+2r}{2}-1}}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1+n_2}{2}}} dF.$$

Let us put, $\frac{n_1}{n_2}F = y \quad \therefore F = \frac{n_2}{n_1}y$, or, $dF = \frac{n_2}{n_1}dy$; $0 \le y \le \infty$.

$$\text{Then, } \mu'_r = \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \int_0^\infty \frac{\left(\frac{n_2}{n_1}y\right)^{\frac{n_1+2r}{2}-1}}{(1+y)^{\frac{n_1+n_2}{2}}} \frac{n_2}{n_1} dy$$

$$= \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \left(\frac{n_2}{n_1}\right)^{\frac{n_1+2r}{2}}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \int_0^{\infty} \frac{y^{\frac{n_1+2r}{2}-1}}{(1+y)^{\frac{n_1+n_2}{2}}} dy.$$

$$= \frac{\left(\frac{n_2}{n_1}\right)^r}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \beta\left(\frac{n_1+2r}{2}, \frac{n_2-2r}{2}\right)$$

$$= \frac{\left(\frac{n_2}{n_1}\right)^r \Gamma\left(\frac{n_1+2r}{2}\right) \Gamma\left(\frac{n_2-2r}{2}\right)}{\Gamma\frac{n_1}{2} \Gamma\frac{n_2}{2}}$$

$$= \frac{\left(\frac{n_2}{n_1}\right)^r \left(\frac{n_1+2r}{2}-1\right)! \left(\frac{n_2-2r}{2}-1\right)!}{\left(\frac{n_1}{2}-1\right)! \left(\frac{n_2}{2}-1\right)!}, r < \frac{n_2}{2}$$

$$= \left(\frac{n_2}{n_1}\right)^r \frac{\left(\frac{n_1}{2}+r-1\right)! \left(\frac{n_2}{2}-r-1\right)!}{\left(\frac{n_1}{2}-1\right)! \left(\frac{n_2}{2}-1\right)!}$$

Thus $\mu_1' = \frac{n_2}{n_2-1}$, $n_2 > 2$. This is independent of n_1 and is always greater than 1.

$$\mu_2' = \frac{n_2^2 (n_1+2)}{n_1 (n_2-2)(n_2-4)}, n_2 > 4.$$

$$\therefore \mu_2 = \frac{n_2^2}{(n_2-2)^2} \left\{\frac{2(n_1+n_2-2)}{n_1(n_2-4)}\right\}.$$

Similarly $\mu_3' = \frac{n_2^3(n_1+2)(n_1+4)}{n_1^2 (n_2-2)(n_2-4)(n_2-6)}$; $n_2 > 6$.

and $\mu_4' = \frac{n_2^4(n_1+2)(n_1+4)(n_1+6)}{n_1^3(n_2-2)(n_2-4)(n_2-6)(n_2-8)}$; $n_2 > 8$.

The corrected μ_3 and μ_4 can be calculated by the formulae,

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

Thus it is seen that moments of F-distribution depends on only n_1 and n_2. The curve is J-shaped if $n_2 < 2$ and positively skew for $n_2 > 2$. The frequency curve for $n_2 > 2$ is shown below :

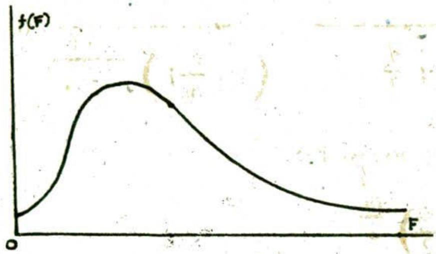


Fig 10.3 F-probability Curve.

The mode of the distribution can be obtained at $F = \frac{n_1 - 2}{n_1} \cdot \frac{n_2}{n_2 + 2}$.

Thus mode of F-distribution is always less than 1.

Inter-relationship Between χ^2, t and F-distribution.

Theorem 10.7 The square of t variate with n d. f. is distributed as F with 1 and n. d. f.

Proof : Let us put $F = t^2$, $n_1 = 1$ and $n_2 = n$ then the distribution of F with n_1 and n_2 d. f. can be written as

$$dF(F) = \frac{\left(\frac{1}{n}\right)^{\frac{1}{2}} (t^2)^{\frac{1-2}{2}}}{\beta\left(\frac{1}{2}, \frac{n}{2}\right)\left(1 + \frac{t^2}{n}\right)^{\frac{1+n}{2}}} d(t^2).$$

$$= \frac{1}{\sqrt{n}} \frac{1}{\beta\left(\frac{1}{2}, \frac{n}{2}\right)} \frac{dt}{\left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}}$$

which is t-distribution with n. d. f.

Theorem 10.8 When n_2 tends to infinity. n_1F. tends to be distributed as a χ^2 with n_1 d. f.

Proof : We know,

$$f(F) = \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \Gamma\frac{n_1+n_2}{2}}{\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} \; \frac{F^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1+n_2}{2}}}; \; 0 \le F \le \infty.$$

In the limit as $n_2 \to \infty$, we have

$$\frac{\Gamma\frac{n_1+n_2}{2}}{n_2^{\frac{n_1}{2}}\Gamma\frac{n_2}{2}} \to \frac{\left(\frac{n_2}{2}\right)^{\frac{n_1}{2}}}{(n_2)^{\frac{n_1}{2}}} = \frac{1}{2^{\frac{n_1}{2}}}$$

We can find out the above by using Starling's approximation and taking the limit $\frac{\Gamma(n+k)}{\Gamma n} \to n^k$ as $n \to \infty$.

$$\text{Also} \lim_{n_2\to\infty} \left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1}{2}} \lim_{n_2\to\infty} \left(1+\frac{n_1}{n_2}F\right)^{\frac{n_2}{2}}$$

$$= e^{-\frac{n_1F}{2}} = e^{-\frac{\chi^2}{2}}.$$

Hence in the limit, the p. d. f. of $n_1 F = \chi^2$ becomes

$$= \frac{1}{2^{\frac{n_1}{2}}\Gamma\frac{n_1}{2}} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n_1}{2}-1} d\chi^2 ; \; 0 \le \chi^2 \le \infty$$

which is the required p. d. f. of chi-square distribution with n_1 d. f.

10.5 Test of Significance

Test of significance is a statistical procedure to arrive at a conclusion or decision on the basis of samples and to test whether the formulated hypothesis can be accepted or rejected in probability sense. The aim of test of significance is to reject the null hypothesis (defined later).

Statistical Hypothesis : A hypothesis concerning the parameters or the form of the probability distribution which we try to varify on the informations provided by a sample, is called statistical hypothesis.

Parametric and Non-parametric Hypothesis : When the hypothesis concerning the parameters of the distribution, provided the form of the distribution is known, is called parametric hypothesis. While the hypothesis regarding the form of the distribution with specified or unspecified parameters, is called non-parametric hypothesis. For example, the hypothesis regarding the population mean and variance of a normal distribution may be considered as parametric hypothesis and the hypothesis that the sample has been obtained from binomial distribution with known or unknown probability of success may be considered as non-parametric hypothesis.

Null Hypothesis and Alternative Hypothesis : The hypothesis which we are going to test for possible rejection under the assumption that it is true is called the null hypothesis, usually denoted by H_o and each of all possible hypothesis other then H_o is called alternative hypothesis, denoted by H_1.

For example, if $H_0 : \mu_1 = \mu_2$

then i) $H_1 : \mu_1 < \mu_2$. ii) $H_1 : \mu_1 > \mu_2$ etc. are alternative hypotheses.

Simple and Composite Hypothesis : If the hypothesis specifies all the parameters of the distribution, is called simple hypothesis otherwise it is called composite hypothesis. For example, a normal distribution has two parameters μ and σ^2. The hypothesis $H : \mu = \mu_o$ and $\sigma^2 = \sigma^2_0$ is simple hypothesis while the hypothesis regarding either of these two parameters is composite hypothesis. There may be number of composite hypotheses of the above case.

Error of 1st and 2nd Kind : We may commit two types of errors for making any conclusion on H_0 on the basis of sample. The error of rejecting H_0 (accepting H_1) when it is true is called the error of 1st Kind or Type I error

The error of accepting H_0 when it is false (H_1 is true) is called the error of 2nd Kind or Type II error.

Critical Region and Acceptance Region : Let $x_1, x_2, x_3, \ldots\ldots x_n$ be a sample point designated by X in an n-dimensional sample space. If X falls in the region for which we reject H_0 when it is true then the region is called critical region denoted by ω, say; and if X falls in the rest of the sample space, $\bar{\omega}$ we accept H_0, in that case $\bar{\omega}$ is called the acceptance region.

Level of significance : The probability of Type I error, denoted by α, is called the level of significance, i.e. $P\{X \text{ falls in } \omega/H_0\} = \alpha$

We usually consider 5% and 1% level of significance for testing hypothesis.

Power of the test : Let the probability of Type II error be β i.e.
$P\{X \text{ falls in } \bar{\omega}/H_1\} = \beta$, then $P\{X \text{ talls in } \omega/H_1\} = 1 - \beta$ which is called the power of the test.

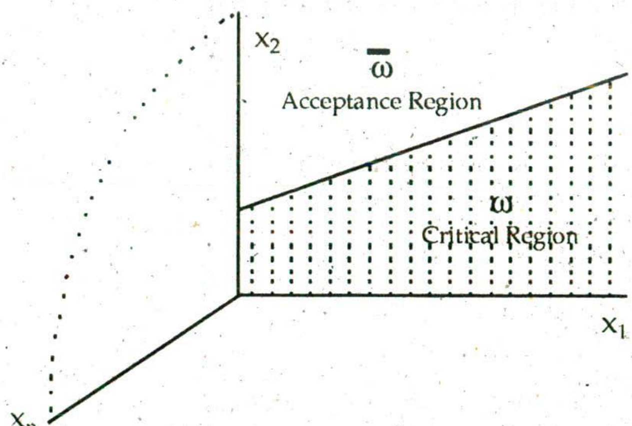


Fig 10.4 Critical and acceptance region.

10.6 Some Important Test of Significance and their Applications

Some of the important tests of significance used mainly in statistics are

1) Normal test. 2) t-test. 3) χ^2-test. 4) F-test.

The description and applications of the above tests are briefly discussed in the next page.

1) **Normal test :** Let u be the statistic whose expected value is E(u), specified by the null hypothesis and its standard error, $\sigma(u)$ is either known or can be estimated from large sample (sample size ≥ 30) then

$$|d| = \frac{u - E(u)}{\sigma(u)} \qquad \text{..........(10.14)}$$

which is distributed normally with mean 0 and variance 1 i. e. d is N (0,1) variate. When u is normal then d is exactly N(0, 1) variate and a normal test can be applied. Again when u is not normally distributed and $\sigma(u)$ is estimated from large sample then d is approximated satisfactorily to normal distribution and in that case also a normal test can be carried out. That is why it is often called a large sample test.

Normal test is usually two-tail test. From normal probability table we get. Prob $[-1.96 \leq d \leq 1.96] = 0.95$ which

implies that, Prob $[|d| \leq 1.96] = 0.95$ also

implies that, Prob $[|d| \geq 1.96] = 1 - 0.95 = 0.05$.

and similarly we can get, Prob $[|d| \geq 2.58] = 1 - 0.99 = 0.01$.

Thus the significant value of $|d|$ at 5% and 1% level of significances are 1.96 and 2.58 respectively. The conclusion regarding the null hypothesis H_0 can be made as follows :

i) If $|d| < 1.96$, the value of $|d|$ is insignificant and H_0 may be accepted.

ii) If $1.96 \leq |d| < 2.58$, the value of $|d|$ is significant and H_0 may be rejected at 5% level of of significance. (A)

iii) If $|d| > 2.58$, the value of $|d|$ is highly significant and H_0 may be rejected.

Uses : This test is used for testing hypothesis regarding means, proportions and correlation co-efficients.

Applications of Normal Test :

(1.a) Test of significance for single mean

Let us suppose that $x_1, x_2, \ldots\ldots x_n$ be a random sample of size n from a normal population with known variance. We want to test the null hypothesis that the population mean is equal to some assigned value say μ_0 i.e. $H_0 : \mu = \mu_0$ (specified value).

The test statistic is

$$|d| = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \qquad \text{.............(10.14a)}$$

which is distributed as N(0, 1) variate.

If σ^2 is not known and the sample size is greater than 30, σ in (10.14a) is replaced by its estimate from the sample. This test is also a normal test.

The conclusion can be made following the principle given in (A).

Example 10.1 A sample of 400 items is drawn from a normal population whose mean is 5 and variance is 4. The sample mean is 4.45. Can the sample be regarded as true random sample drawn from the population?

Solution : Let the null hypothesis be $H_0 : \mu = 5$.

The statistic is $|d| = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{4.45 - 5}{\dfrac{2}{\sqrt{400}}} = 5.5$

which is distributed as N(0, 1) variate.

The calculated value of $|d|$ is greater than 2.58, hence it is highly significant and the hypothesis may be rejected.

(1.b) Test of significance of difference of means

Let $\bar{x}$ be the mean of a random sample of size n_1 from a normal population with mean μ_x and known variance σ_x^2 and let $\bar{y}$ be the mean of an independent random sample of size n_2 from another normal population with mean μ_y and known variance σ_y^2. For testing the null hypothesis, $H_0 : \mu_x = \mu_y$,

the required test statistic is $|d| = \dfrac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_x^2}{n_1} + \dfrac{\sigma_y^2}{n_2}}}$ ....(10.15)

which is distributed as N(0.1) variate

If $\sigma_x^2 = \sigma_y^2 = \sigma^2$, then the test statistic is,

$$|d| = \frac{\overline{x} - \overline{y}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{..............(10.16)}$$

which is also distributed as N(0, 1) variate.

Even if σ_x^2 and σ_y^2 are not known but $n_1 > 30$ and $n_2 > 30$ then σ_x^2 and σ_y^2 are replaced in (10.15) by their estimates s^2_x and s^2_y respectively from the samples where $s_x^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \overline{x})^2$;

$$\overline{x} = \frac{\sum^{n_1} x_i}{n_1} \text{ and } s_y^2 = \frac{1}{n_2} \sum_{j=1}^{n_2} (y_j - \overline{y})^2; \; \overline{y} = \frac{\sum^{n_2} y_i}{n_2}$$

And again, if σ in (10.16) is not known and the samples are large, (n_1, $n_2 > 30$) the estimate of σ is replaced in (10.16). The estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2}$$

If the hypothesis to be tested is that the population means are μ_x and μ_y (some specified values), we can carry out the test of significance as above considering the numerator of the test statistic $|d|$ as $(\overline{x} - \overline{y}) - (\mu_x - \mu_y)$.

The conclusions of the above hypotheses can be done following the principles given in (A).

Example 10.2 The mean yields of two sets of plots and their variability are given below. Test the hypothesis that the difference in the mean yields of the two sets of plots is significant.

	Set of 40 plots	Set of 60 plots
Mean yield/plot -	1258	1243
S. D. per plot -	34	28

Solution : We set up H_0: $\mu_1 = \mu_2$ where μ_i represents the population mean of the ith set.

The test statistic is, $|d| = \dfrac{1258 - 1243}{\sqrt{\dfrac{(34)^2}{40} + \dfrac{(28)^2}{60}}} = \dfrac{15}{\sqrt{41.92}} = 2.3.$ (app)

Since the calculated value of $|d|$ is greater than 1.96, but less than 2.58, it is significant and the hyphothesis may be rejected at 5% level of significance. In such case, further investigations are, advised to get exact conclusion.

(1.c) Test of significance for sample proportion

Let us consider an independent random sample from a binomial population of size n > 30 of which x is the number of individuals which possess certain characteristic, then the observed proportion of the individuals possessing that characteristic is given by $p = \frac{x}{n}$. We are to test the null hypothesis,

$H_0 : \pi = \pi_0$ (a specified value) where π is the population proportion.

The required test statistic is

$$|d| = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1 - \pi_0)}{n}}} \qquad \text{..........(10.17)}$$

which is distributed as N(0,1) variate.

The conclusion can be made following the principles given in (A).

Example 10.3 A random sample of 100 seeds was taken from a large consignment for examination and 12 were found to be defective. Can we accept the suppliers claim that the proportion of bad seeds in the consignment is 0.02 ?

Solution : We set up $H_0 : \pi = 0.02$.

We calculate $p = \frac{12}{100} = 0.12$ and $s.e(p) = \sqrt{\dfrac{\pi_0(1 - \pi_0)}{n}}$

$= \sqrt{\dfrac{0.02 \times 0.98}{100}} = \dfrac{.14}{10} = .014$

The required test Statistic, $|d| = \dfrac{.12 - 0.2}{.014} = 7.1$ (app)

The calculated value of $|d|$ is highly significant and therefore the hypothesis may be rejected.

(1.d) Test of significance for difference of proportions

Let us suppose that we have two independent samples of sizes n_1 and n_2 (n_1, n_2 > 30) obtained from two seperate binomial populations of which x_1 and x_2 are the number of individuals possessing certain characteristic. The observed proportions of two samples being p_1 and p_2 respectively. We are to test the hypothesis that the two samples have been drawn from same binomial population. i.e. $H_0 : \pi_1 = \pi_2$,

We calculate $p_1 = \frac{x_1}{n_1}$ and $p_2 = \frac{x_2}{n_2}$.

The combined proportion of two samples is, $p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$ and $q = 1 - p$.

The required test Statistic is $$|d| = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \text{.........(10.18)}$$

which is distributed as N(0,1) variate.

The conclusion can be made as according to the princples given in (A).

If the hypothesis is to test whether the population proportions are π_1 and π_2 (some specified values) the test statistic becomes

$$|d| \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}} \qquad \text{.....(10.19)}$$

which is also distributed as N(0,1) variate and the conclusion can be made according to the principle given in (A).

Examples 10.2 In a year there are 956 births in town A of which 52.5% were males while in towns A and B combined, this proportion in a total of 1406 births was 0.495. Is there any significant difference in the proportion of male births in the two towns?

Solution : We set up the null hypothesis, $H_0 : \pi_1 = \pi_2$ i.e. there is no significant difference in the proportion of male births in the two towns.

We know, $n_1 = 956$ and $n_1 + n_2 = 1406$ $\qquad \therefore n_2 = 450$.

$p_1 = 0.525$ and the combined proportion is

$$\frac{956 \times 0.525 + 450 \times p_2}{956 + 450} = 0.496 \qquad \therefore \quad p_2 = 0.432.$$

The test statistic $|d| = \dfrac{0.525 - 0.432}{\sqrt{0.496 \times 0.504 \left(\dfrac{1}{956} + \dfrac{1}{450}\right)}}$

$$= \frac{93}{28.59} = 3.25 \text{ (app)}$$

The calculated value of $|d|$ is greater than 2.58, hence it is highly significant and the H_0 may be rejected i.e. there is evidence that there is significant difference in the proportions of male births in towns A and B.

(1.c) Test of significance of specified value of population correlation co-efficient

Let us suppose that we have a random sample of n pairs of values from a bivariate normal population. The calculated value of the correlation co-efficient is, r, say. For testing the null hypothesis that the population correlation co-efficient is ρ_0, a specified value i.e. H_0: $\rho = \rho_0$ (a specified value), we need a transformation known as Fisher's z transformation for correlation co-efficient available in Table No. 14, page-139 ; Vide Biometrika Tables for Statisticians edited by E. S. Pearson and O. H. hartley.

This is defined by, $z = \frac{1}{2} \log_e \frac{1 + r}{1 - r}$. This transformation is useful for the following reasons namely the distribution of r is far form normal and changes as ρ, the population correlation co-efficient changes. But the distribution of z is approximately normally distributed with mean, $m = \frac{1}{2} \log_e \frac{1 + \rho_0}{1 - \rho_0}$ and variance $= \frac{1}{n - 3}$.

The test statistic is $|d| = \dfrac{(z - m)}{\sqrt{\dfrac{1}{(n-3)}}} = (z - m)\sqrt{(n - 3)}$ ....(10.20)

which is distributed as N(0,1) variate. The conclusion can be made following the principles given in (A).

Example 10.5 In a random sample of 28 pairs of values from a bivariate normal population, the correlation co-efficient was found 0.7. Is this value consistent with the assumption that the correlation co-efficient in the population is 0.5?

Solution : We set up the null hypothesis, $H_0 : \rho = 0.5$.

From z-transformation, we have, $r = 0.7$; $z = 0.87$ $\rho = 0.5$; $m = 0.55$

The test statistic is $|d| = \dfrac{0.87 - 0.55}{\sqrt{\dfrac{1}{(28-3)}}} = 1.6$

The calculated value of $|d|$ is less than 1.96, hence it is insignificant and the hypothesis may be accepted i.e. the population correlation co-efficient is 0.5.

(1.f) Test of significance of the difference of correlation co-efficients

Let r_1 and r_2 be the sample correlation co-efficients obtained from two independent random samples of sizes n_1 and n_2 respectively obtained from two seperate bivariate normal populations. We are to test the hypothesis that the samples are drawn from two different populations with same correlation co-efficient or from same population.

Let us obtain the values of z_1 and z_2 from the Table No. 14, Page 139 ; Vide Biometrika Tables for Statisticians Edited by E. S. Pearson and O. H. Hartley. We know, $z_1 = \frac{1}{2}\log_e \frac{1+r_1}{1-r_1}$ and $z_2 = \log_e \frac{1+r_2}{1-r_2}$

Then under $H_0 : \rho_1 = \rho_2$; $(z_1 - z_2)$ is approximately normally distributed with zero mean and variance $\left[\dfrac{1}{(n_1-3)} + \dfrac{1}{(n_2-3)}\right]$.

The required test statistic is $|d| = \dfrac{(z_1 - z_2)}{\sqrt{\dfrac{1}{n_1-3} + \dfrac{1}{n_2-3}}}$ ....(10.21)

which is distributed as N(0,1) variate.

The conclusion can be drawn as given the principles in (A).

Example 10.6 The correlation co-efficients obtained from samples of sizes 20 and 32 are 0.47 and 0.68 respectively. Test the significance of the difference between these co-efficients.

Solution : We set up H_o : $\rho_1 = \rho_2$.
Here, $n_1 = 20$, $r_1 = 0.47$
and $n_2 = 32$, $r_2 = 0.68$.

From z-transformation, we have, $z_1 = 0.51$, and $z_2 = 0.83$.

The required test statistic is, $|d| = \dfrac{0.51 - 0.83}{\sqrt{\dfrac{1}{17} + \dfrac{1}{29}}} = \dfrac{0.32}{.305} = 1.05$ (app)

Since the calculated value of $|d|$ is less than 1.96, it is insignificant and the hypothesis may be accepted.

2) t-test : In normal test, we assume that $\sigma(u)$ in (10.14) is either known or can be estimated from a large sample ($n > 30$). We may have to face some situations where the sample sizes are not large enough and also the $\sigma(u)$ is not known. In such case, the estimate of $\sigma(u)$ can be obtained and the test statistic becomes

$$t = \frac{u - E(u)}{\text{estimated } \sigma(u)} \qquad \text{.........(10.22)}$$

which is distributed as Student's t with δ d. f. where δ is less than n, the sample size, mainly depends on the d. f. of the estimated $\sigma(u)$.

When δ is large, t-test becomes normal test, therefore, t-test is small sample test and can be considered as a special case of normal test. Like normal test, t-tests are two tail tests. The theoretical or tabulated value of t for different d. f. as well as different levels of significance are given in Table No. III, Page-46, Vide Statistical Tables for Biological, Agricultural and Medical Research.

The conclusion can be made as below :

i) If the calculated value of $|t|$ with δ d. f. (say), is smaller than the tabulated value of t with same d. f. at 5% level of significance then the value of $|t|$ is insignificant and the null hypothesis may be accepted.

ii) If the calculated value of $|t|$ with δ d. f. (say) is greater than the tabulated value of t with same d. f. at 5% level of significance but smaller than the value of t with same d. f. at 1% level of significance then the value of $|t|$ is significant and the null hypothesis may be rejected at 5% level of significane. (B)

iii) If the calculated value of $|t|$ with δ d. f. is greater than the tabulated value of t with same d. f. at 1% level of significance then the value of $|t|$ is highly significant and the null hypothesis may be rejected.

Uses : t-test is used to test the null hypothesis regarding means, correlation co-efficients and regression co-efficients.

Applications of t-test

(2.a) Test of significance of single mean

Let us suppose that $x_1, x_2, \ldots, x_n$ be a random sample of size n ($n < 30$), drawn from a normal population with known mean and unknown variance. We are to test the null hypothesis H_0, that the sample has been drawn from a population with mean μ_0 (a specified value) i.e. $H_0 : \mu = \mu_0$ (a specified value).

Since population variance σ^2 is not known the unbiased estimate of it is given by, $s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2$.

The required test statistic is $|t| = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$ ....(10.23)

which is distributed as Student's t with (n -1) d.f.

The conclusion can be made following the principles given in (B).

Example 10.2 Ten plots of same area are chosen at random and the yield of a certain paddy variety are recorded in kg., they are 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71. In the light of above data can you suggest that the population mean production of that paddy variety is 66 kg. for same area?

Solution : We set up $H_0 : \mu = 66$.

Here $\overline{x} = 67.8$ kg. and $s = \sqrt{\frac{1}{9}\Sigma(x_i - \overline{x})^2} = 3.011$ k.g.

The test statistic, $t = \dfrac{67.8-66}{3.011/\sqrt{10}} = 1.89$ (app) with 9 d.f.

The calculated value of t with 9. d. f. is seen to be smaller than the tabulated value of t at 5% level of significance i.e. $t_{0.05} = 2.26$, with 9 d. f. Hence the calculated value is insignificant and the hypothesis may be accepted.

(2.b) Test of significance of difference of means

Let $\overline{x}$ be the mean of a random sample of size $n_1 < 30$ from a normal population with known mean μ_x and unknown variance and let $\overline{y}$ be the mean of another independent random sample of size $n_2 < 30$ from another normal population with known mean μ_y and unknown variances. The variance of the two populations are assumed to be equal. For testing $H_0: \mu_x = \mu_y$ (some specified mean values)

The required test statistic is, $$|t| = \frac{(\overline{x} - \overline{y}) - (\mu_x - \mu_y)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{..........(10.24)}$$

which is distributed as t with $(n_1 + n_2 - 2)$ d. f.

where $\overline{x} = \frac{\sum_{}^{n_1} x_i}{n_1}$; $\overline{y} = \frac{\sum_{}^{n_2} y_j}{n_2}$ and

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(x_i - \overline{x})^2 + \sum_{j=1}^{n_2}(y_j - \overline{y})^2\right] \quad \text{......(10.25)}$$

When H_0 : The two population means are same i. e. $\mu_x = \mu_y$,

$$|t| = \frac{\overline{x} - \overline{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{..........(10.26)}$$

which is distributed as t with $(n_1 + n_2 - 2)$ d. f. and s in defined as in (10.25). when $n_1 = n_2 = n$, the statistic becomes

$$|t| = \frac{\overline{x} - \overline{y}}{s\sqrt{\frac{2}{n}}} \quad \text{..........(10.27)}$$

which is t with $(2n - 2)$ d. f.

The conculusion can be made as given the principles in (B).

Reamrk : For testing above hypotheses given in (10.24) and (10.26) it is desirable to test the equality of population variances by applying F-test (given latter on). If the variances donot come out to be equal, the following test is to be performed.

When population variances are not equal the required test statistic under H_0 is given by,

$$t' = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}} \qquad \text{..........(10.28)}$$

t' given in (10.28) is not a student's t. The tabulated value of t at $\alpha\%$ level of significance can be obtained from the following formula,

$$t'_\alpha = \frac{\frac{s_x^2 t_1}{n_1} + \frac{s_y^2 t_2}{n_2}}{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}$$

where t_1 and t_2 are Student's t with $(n_1 - 1)$ and $(n_2 - 1)$ d. f. respectively at α % level of significance.

If $n_1 = n_2$, then $t_1 = t_2 = t$ say, which implies that $t'_\alpha = t$.

When the null hypothesis indicates the specified value of the population means, say μ_x and μ_y, the test statistic becomes,

$$t' = \frac{(\overline{x} - \overline{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}} \qquad \text{..........(10.29)}$$

The conclusion can be drawn as given the principles in (B).

Example 10.8 The following data represent the yield in bushels of corn on ten subdivisions of equal areas of two agricultural plots in which plot I was a central plot treated the same as plot-II except for the amount of phosphorus applied as a fertiliser :

Plot-I : 6.2, 5.7, 6.5, 6.0, 6.3, 5.8, 5.7, 6.0, 6.0, 5.8
Plot-II : 5.6, 5.9, 5.6, 5.7, 5.8, 5.7, 6.0, 5.5, 5.7, 5.5.

Is there significant difference between the yields on the two plots. using the difference between their means as a criterion of judgment?

Solution : Let x and y be variable for plot - I and plot - II respectively.

We calculate $\overline{x} = \frac{\Sigma x}{10} = \frac{60}{10} = 6.$ $\qquad \overline{y} = \frac{\Sigma y}{10} = \frac{57}{10} = 5.7.$

$\Sigma(x_i - \overline{x})^2 = 0.64$ and $\Sigma(y_j - \overline{y})^2 = 0.24.$

$\therefore$ Pooled variance, $s^2 = \frac{0.64 + 0.24}{10 + 10 - 2} = \frac{0.88}{18} = 0.049$

The required test statistic for testing $H_0 : \mu_x = \mu_y$ is

$$t = \frac{0.3}{\sqrt{0.049\left(\frac{1}{10}+\frac{1}{10}\right)}} = 3.03 \text{ (app.) with 18 d. f.}$$

Since the calculated value of t with 18 d. f. is greater than the tabulated value of t with 18 d.f. at 5% level of significance, the value is significant and the hypothesis may be rejected at 5% level of significance.

(2.c) Test of significance for difference of means from correlated populations

Let us consider the situation where the sample sizes are same i. e. $n_1 = n_2 = n$. The two samples are not independent and the samples are paired together. The situation may arise for the case where for avoiding extraneous influence we consider a plot of land which is equally divided and two types of paddy varities say, Irri and Boro are sown, thus giving us a pair of observations of yields of Irri and Boro. Let us consider such n pairs of observation. Now we are to test the null hypothesis whether the sample means differ significantly or not.

Let x_i and y_i (i = 1, 2,........,n) be the yields on the ith plot and $d_i = x_i - y_i$. We set up the null hypothesis, $H_0 : \mu_d = \mu_x - \mu_y = 0$.

It is assumed that $d_1, d_2, \ldots, d_n$ constitute a random sample from a normal population with mean μ_d and variance σ_d^2 (unknown). The required test statistic is.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \text{............(10.30)}$$

which is distributed as t with (n -1) d. f., $\bar{d}$ is the mean of d_i's & s^2_d is the sample variance of d_i's based on (n -1) d. f.

The conclusion can be made as given the principles in (B).

Example 10.9 The following table shows the mean number of bacterial colonies per plate obtainable by four slightly different methods from soil samples taken at 4 P. M. and 8 P. M. respectively.

	Methods	A	B	C	D
Time	4 P.M.	29.75	27.50	30.25	27.80
	8 P.M.	39.20	40.60	36.20	42.40

Are there significantly more bacteria at 8 P. M. than at 4 P.M.?

Solution : Calculations of mean and standard deviation :

Methods	4 P.M. (x)	8 P.M. (y)	$d = y - x$	$d - \bar{d}$	$(d - \bar{d})^2$
A	29.75	39.20	9.45	-1.325	1.756
B	27.50	40.60	13.10	2.325	5.406
C	30.25	36.20	5.95	-4.825	23.281
D	27.80	42.40	14.60	3.825	14.631.

We have, $\bar{d} = \frac{\Sigma d_i}{n} = \frac{43.10}{4} = 10.775.$ and $s^2_d = \frac{\Sigma(d_i - \bar{d})^2}{n-1} = 15.025$

The test statistic for testing $H_0 : \mu_d = \mu_x - \mu_y = 0$ is

$$t = \frac{10.775}{\sqrt{15.025/4}} = 5.56 \text{ (app) with } 4 - 1 = 3 \text{ d. f.}$$

Since the calculated value of t with 3 d. f. is greater than the tabulated value of t with 3 d. f. at 5% level of significance, the calculated value is significant and the hypothesis may be rejected at 5% level of significance.

(2.d) Test of significance of an observed correlation co-efficient

Let us suppose that r be the correlation co-efficient from a sample of size n from a bivariate normal population. We are to test the null hypothesis that the population correlation co-efficient is zero, i. e. $H_0 : \rho = 0$.

The required test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{..........(10.31)}$$

which is distributed as t with (n -2) d. f.

The conclusion can be drawn as given the principles in (B).

Remarks :

(1) The same test statistic can be used if we want to test the null hypothesis $H_0 : \beta = 0$. where β is the regression co-efficient of y on x. Here the usual assumption is that x is an $N(\mu_x, \sigma^2)$ variate and y is a fixed variate for $\beta = 0$.

(2) The same test statistic is used for testing the null hypothesis regarding the population rank correlation co-efficient is equal to zero. In this case, in the test statistic r is replaced by R, the sample rank correlation

Example 10.10 A random sample of 18 pairs from a bivariate normal population showed a correlation co-efficient 0.3. Is this value significant of correlation in the population?

Solution : We set up the null hypothesis, $H_0 : \rho = 0$.

The test statistic is, $|t| = \dfrac{0.3\sqrt{18-2}}{\sqrt{1-0.09}} = 1.26$ (app) with 16 d. f.

The calculated value of t with 16 d. f. is seen to be smaller than the tabulated value of t with same d. f. at 5% level of significance. Hence the calculated value of $|t|$ is insignificant and the hypothesis may be accepted.

(2.e) Test of significance of an observed regression co-efficient

Let us suppose that (x_i, y_i), $(i = 1, 2, \ldots\ldots, n)$, be a random sample of size n of which x_i's are random and y_i's are fixed. We are to test the null hypothesis that the regression co-efficient of y on x is β_0 (a specified value), i.e. $H_0 : \beta = \beta_0$ (a specified value).

The line of regression of y on x is $y - \overline{y} = b(x - \overline{x})$ .....(10.32)

where $b = \dfrac{S.P.(xy)}{S.S.(x)}$. The estimate of y for a given value x_i (say) of x as given by the line (10.32) is $\hat{y}_i = \overline{y} + b(x_i - \overline{x})$.

The required test statistic is

$$|t| = (b - \beta_0)\left[\frac{(n-2)\Sigma(xi - \overline{x})^2}{\Sigma(yi - \hat{y}_i)^2}\right]^{\frac{1}{2}} \qquad .....(10.33)$$

which is distributed as t with (n - 2) d.f.

The conclusion can be made as given the principles in (B).

Remark : Sometimes we may want to test the hypothesis that α, the constant term or intercept of the regression equation takes a particular value say, α_0 i.e. $H_0 : \alpha = \alpha_0$ (a specified value).

From the regression equation $y_i = a + b\,x_i$ the value of 'a' can be obtained by $a = \overline{y} - b\,\overline{x}$ where $b = \dfrac{S.P.(x,y)}{S.S.(x)}$; $\overline{y}$ and $\overline{x}$ are the means of y_i and x_i's respectively.

The test statistic is $|t| = \dfrac{(a - \alpha_0)\sqrt{[n(n-2)\Sigma(x_i - \overline{x})^2]}}{\sqrt{[(\Sigma x_i^2)\,\Sigma(y_i - \overline{y}^2)]}}$ .....(10.34)

which is distributed as t with (n - 2) d. f.

The conclusion can be made as given the principles in (B).

(2.f) Test of significance of difference of regression co-efficients

Let us suppose that we have b_1 and b_2, two estimates of same regression co-efficient in two different times or samples taken by two investigators. We are interested to test the null hypothesis $H_0 : \beta_1 = \beta_2$ i. e. the two samples have been drawn from the same population.

The test statistic is

$$|t| = \frac{b_1 - b_2}{s\sqrt{\dfrac{1}{\sum\limits_{i=1}(x_{1i} - \overline{x}_1)^2} + \dfrac{1}{\sum\limits_{j=1}(x_{2j} - \overline{x}_2)^2}}} \qquad .....(10.35)$$

which is distributed as t with $(n_1 + n_2 - 4)$ d. f. where

$s^2 = \dfrac{(n_1 - 2)s_1^2 + (n_2 - 2)s_2^2}{n_1 + n_2 - 4}$ of which $s_1^2 = \dfrac{\Sigma(y_{1i} - \hat{y}_{1i})^2}{n_1 - 2}$

and $s_2^2 = \dfrac{\Sigma(y_{2j} - \hat{y}_{2j})^2}{n_2 - 2}$; n_1 and n_2 are the sizes of two different samples.

The conclusion can be made as given the principles in (B).

(2.g) Testing significance of an observed partial correlation co-efficient

Let $r_{12.34.........(k+2)}$ be the partial correlation co-efficient of order k, calculated from a sample of size n from a multivariate normal population, we want to test the null hypothesis that the population partial correlation co-efficient is zero i. e. $H_0 : \rho_{12.34.......(k+2)} = 0$.

The required test statistic is, $t = \dfrac{r_{12.34.......}\sqrt{n-k-2}}{\sqrt{1 - r^2_{12.34.......}}}$ ....(10.36)

which is distributed as t with (n - k - 2) d. f.

The conclusion can be made as given the principles in (B).

Example 10.11. Partial correlation co-efficient $r_{12.34} = 0.5$ is obtained from a sample of size 20 from a 4-variate normal population. Test its significance.

Solution : We set up the null hypothesis $H_0 : \rho_{12.34} = 0$.

Here $r_{12.34} = 0.5$, $n = 20$, $k = 2$.

The required test statistic, $t = \dfrac{0.5\sqrt{16}}{\sqrt{1 - .25}} = \dfrac{2}{\sqrt{.75}} = 2.31$ (app) with 16 d. f.

The calculated value of t is greater than the tabulated value of t at 5% level of significance. Hence the calculated value of t is significant and the null hypothesis may be rejected at 5% level of significance.

3) χ^2-test : χ^2-test is mainly used to test the hypothesis which specifies the nature of one or more distributions. We know the mathematical form of the distribution, hypothesis regarding the sample that has been drawn from the distribution is tested by χ^2-statistic. We may be interested to test whether two or more distributions are identical. It also tests the independence of two or more attributes. For testing the above hypotheses, we used to compare an observed set of frequencies with a corresponding set of frequencies that are expected under the null hypothesis. Let O_i ($i = 1, 2, \ldots, k$) denote the observed frequencies and E_i ($i = 1, 2, \ldots, k$) denote the expected frequencies then the test statistic, χ^2 is defined as,

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n \qquad \ldots\ldots\ldots\ldots(10.37)$$

where $n = \sum^{k} E_i = \sum^{k} O_i$; which is distributed as χ^2 with $(k - p)$ d. f.

where p is the number of independent restrictions imposed for the calculation of the set of expected frequencies. The d. f. corresponding to each χ^2-test will be specified independently in every case. The above test statistic is an approximation under null hypothesis and is fairly good when the expected frequencies are greater than or equal to 5. For values, less than 5, the modifications are given in the appropriate cases.

Uses : χ^2-test is also used for testing significance of variance, proportions and correlation co-efficients.

The theoretical or tabulated value of χ^2 with different d. f. as well as different levels of significance are given in Table No. IV, Page-47, Vide Statistical Table for Biological, Agricultural and Medical Research.

The conclusion can be drawn as below :

i) If the calculated value of χ^2 with δ d. f (say) is smaller than the tabulated value of χ^2 with same d. f. at 5% level of significance, then the calculated value of χ^2 is insignificant and the null hypothesis may be accepted.

ii) If the calculated value of χ^2 with δ d. f. (say) is greater than the tabulated value of χ^2 with same d. f. at 5% level of significance but smaller than the tabulated value of χ^2 with same d. f. at 1% level of significance then the calculated value of χ^2 is significant and the hypothesis may be rejected at 5% level of significance.

iii) If the calculated value of χ^2 with δ d. f. (say) is greater than the tabulated value of χ^2 with same d. f. at 1% level of significance then the calculated value of χ^2 is highly significant and the null hypothesis may be rejected.

(C)

Applications of χ^2 test

(3.a) χ^2-test for testing goodness of fit : Let us suppose that we are given a sample and the problem is to test the hypothesis that the samples has been drawn from a particular population with some specified or unspecified values of parameters. The sample can be arranged in the frequency distribution. Corresponding to every value of the observed frequencies we can have expected frequencies obtained from the knowledge of the population. Now, if the deviation of the observed frequencies and the expected frequencies are small, we can easily infer that the deviations are due to sampling fluctuation and the sample may be considered to be drawn from that specified population. On the other hand, larger value of the deviations indicate that the given sample could not have possibly come from the population mentioned.

If O_i ($i = 1, 2, \ldots, k$) be a set of observed frequencies and E_i be the corresponding set of expected frequencies, then for large n, $n = \sum^{k} O_i = \sum^{k} E_1$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n \qquad \ldots\ldots\ldots\ldots (10.38)$$

which follows χ^2-distribution with (k - 1) (for specified set of parameters) or (k - b - 1) (for b unspecified parameters) d. f. This test was given by Karl Pearson in 1900. The conclusion can be drawn as given the principles in (C).

Conditions for the validity of χ^2-test for goodness of fit :

1) The sample observation should be independent.

2) The constraint on the cell frequencies is $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$,

3) n, the total frequency should be reasonably large, say, greater than 50.

4) No expected frequency should be less than 5. If any expected frequency is less than 5, then for the application of χ^2-test it is to be pooled with the preceeding or succeeding frequency so that the pooled frequency is more than 5 and finally an adjustment for the loss of d. f. is necessary.

Example 10.12 Test the goodness of fit of the data given in Example 8.2

Solution : We have calculated the expected frequencies in the solution of the Example 8.2. Therefore, we can furnish the required table as follows:

(i) H_0 : The sample has been obtained from a binomial distribution with $p = \frac{1}{2}$.

x	Observed Frequency (O)	Expected Frequency (E)	O^2/E
0 1	$\left.\begin{matrix}7\\6\end{matrix}\right\}$ 13	$\left.\begin{matrix}1\\7\end{matrix}\right\}$ 8	21.125
2	19	21	17.190
3	35	35	35.000
4	30	35	25.714
5	27	21	34.714
6 7	$\left.\begin{matrix}7\\1\end{matrix}\right\}$ 8	$\left.\begin{matrix}7\\1\end{matrix}\right\}$ 8	8
Total	128	128	141.743

Therefore, χ^2 = 141.743 - 128 = 13.743 (app) with 6 - 1 = 5 d. f.

The tabulated value of χ^2 with 5 d. f. at 5% level of significance is 11.07. Our calculated value is 13.743 which is greater than the tabulated value. So the calculated value is significant and the hypothesis may be rejected.

(ii) H_0 : The sample has been obtained from a binomial distribution with unknown p.

x	Observed Frequency (O)	Expected Frequency (E)	O^2/E
0 1	$\left.\begin{matrix}7\\6\end{matrix}\right\}13$	$\left.\begin{matrix}1\\8\end{matrix}\right\}9$	18.778
2	19	23	15.696
3	35	36	34.028
4	30	34	26.471
5	27	19	38.368
6 7	$\left.\begin{matrix}7\\1\end{matrix}\right\}8$	$\left.\begin{matrix}6\\1\end{matrix}\right\}7$	9.143
Total	128	128	142.484

Therefore, $\chi^2 = 142.484 - 128 = 14.484$ (app) with $6 - 1 - 1 = 4$ d. f.

The tabulated value of χ^2 with 4 d. f. at 1% level of significance is 13.277. Our calculated value is 14.484, which is greater then the tabulated value. So the calculated value is highly significant and the hypothesis may be rejected.

(3.b) χ^2-test for testing independence of attributes

We can classify the sample observations according to more than one attributes. Thus an element of the sample, say student may be classified as "dull headed" or " Mediocre" or the " best one" according to the attribute'intelligence' and then be classified as 'male' or 'female' according to the attribute 'sex'. Data arranged in the form of above classes may be termed as contingency table. Here again the compatibility of the observed and the expected frequencies has to be tested in testing the independence of attributes in the contingency table. In contingency table the values of the variables are generally qualitative whereas in correlation table the variables are quantitative. The observations in the cells represent the frequencies in both the cases.

Contingency table and calculation of χ^2 for testing independence of attributes

Let the data be classified into t-classes, $A_1, A_2, \ldots\ldots A_t$ according to attribute A and into r classes $B_1, B_2, \ldots\ldots, B_r$ according to attribute B. Let O_{ij} denote the observed frequency of the cell belonging to ith class of A (i = 1, 2, 3,........t) and j th class of B (j = 1, 2,......, r). Let $O_{i.}$ and $O_{.j}$ denote the totals of all the frequency belonging to ith class of A and jth class of B respectively. The data can be depicted in a t x r contingency table as below :

A / B	A_1	A_2	....	A_i	....	A_t	Total
B_1	O_{11}	O_{21}	....	O_{i1}	....	O_{t1}	$O_{.1}$
B_2	O_{12}	O_{22}	....	O_{i2}	....	O_{t2}	$O_{.2}$
|	|	|		|		|	|
B_j	O_{1j}	O_{2j}	....	O_{ij}	....	O_{tj}	$O_{.j}$
|	|	|		|		|	|
B_r	O_{1r}	O_{2r}	....	O_{ir}	....	O_{tr}	$O_{.r}$
Total	$O_{1.}$	$O_{2.}$	....	$O_{i.}$	....	$O_{t.}$	n

Here we are to test the hypothesis that the attributes A and B from which the sample of size n has been drawn are independent.

Let P_{ij} denote the probability that an element be chosen at random will be the ith class of A and jth class of B. $P_{i.}$ and $P_{.j}$ are the marginal probabilities for the ith class of the attribute A and jth class of the attribute B respectively. Under the null hypothesis i.e. the two attributes A and B are independent we have, $P_{ij} = P_{i.} \times P_{.j}$ and $\sum_i P_{i.} = \sum_j P_{.j} = 1$.

We know that $P_{i.} = \frac{O_{i.}}{n}$ and $P_{.j} = \frac{O_{.j}}{n}$ and also we know that the expected cell frequencies E_{ij} (i = 1, 2,......, t ; j = 1, 2,......., r) for the ith class of the attribute A and jth class of the attribute B can be written as,

$$E_{ij} = nP_{ij} = nP_{i.} \times P_{.j}$$

$$= n \times \frac{O_{i.}}{n} \times \frac{O_{.j}}{n} = \frac{O_{i.} \times O_{.j}}{n}.$$

Thus the expected cell frequency E_{ij} is equal to the product of the marginal totals of the ith class of the attribute A and jth class of the attribute B devided by the total number of the observations in the sample. The test statistic used to test the hypothesis is, $\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$.

$$= \sum_i \sum_j \frac{O^2_{ij}}{E_{ij}} - n \qquad \text{.............(10.39)}$$

which is approximately distributed as χ^2 with (t - 1) (r - 1) d. f.

Since there are (r - 1) row totals and (t - 1) column totals which are independent in a t X r contingency table. Therefore, the d. f. in a t X r contingency table is tr - 1- {(r - 1) + (t - 1)} = (t - 1) (r - 1). The conclusion can be drawn as given the principles in (C).

Example 10.13 Two investigators draw samples from the same town in order to estimate the number of persons falling in the income groups - 'poor', 'middle class', 'well to-do' (The limits of the group are defined in terms of money and are the same for both investigators). Their results are given in Table-10.1.

Table -10.1

Investigators	Income-group Poor	 Middle-class	 Well to-do	Total
A	140	100	15	255
B	140	50	20	210
Total	280	150	35	465

Show that the sampling techniques of the investigators are independent on the economic conditions of the families.

Solution : We set up the null hypothesis that the two attributes, sampling techniques of the investigators and economic conditions of the families are independent.

We know, under the hypothesis, the expected cell frequencies are $E_{ij} = \frac{O_{i.} \times O_{.j}}{n}$. Now we prepare a table of expected cell frequecies.

Table -10.2

Investigators	Income-group Poor	 Middle-class	 Well to-do	Total
A	154	82	19	255
B	126	68	16	210
Total	280	150	35	465

$$\therefore \chi^2 = \frac{(140-154)^2}{154} + \frac{(100-82)^2}{82} + \frac{(15-19)^2}{19} + \frac{(140-126)^2}{126}$$

$$+ \frac{(50-68)^2}{68} + \frac{(20-16)^2}{16} = 13.387 \text{ (app) with } (3-1)(2-1) = 2 \text{ d. f.}$$

The tabulated value of χ^2 with 2 d. f. at 1% of significance is 9.21, which is smaller than the calculated value of χ^2. Hence the calculated value is highly singnificant and the hypothesis may be rejected.

Example10.14 For the 2 x 2 contingency table whose cell frequencies are :

a	b
c	d

show that the value o χ^2 for testing independence is given by

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \text{ where } n = a+b+c+d.$$

Solution : The cont ingency table with marginal totals is as follows :

			Total
	a	b	a + b
	c	d	c + d
Total	a + c	b + d	a + b + c + d = n

Under the hypothesis of independence of attributes,

$$E(a) = \frac{(a+b)(a+c)}{n}. \qquad E(b) = \frac{(a+b)(b+d)}{n}.$$

$$E(c) = \frac{(a+c)(c+d)}{n}. \qquad \text{and } E(d) = \frac{(b+d)(c+d)}{n}.$$

$$\therefore \chi^2 = \frac{[a - E(a)]^2}{E(a)} + \frac{[b - E(b)]^2}{E(b)} + \frac{[c - E(c)]^2}{E(c)} + \frac{[d - E(d)]^2}{E(d)}.$$

$$\text{Now, } \frac{[a - E(a)]^2}{E(a)} = \frac{\left[a - \frac{(a+b)(a+c)}{(a+b+c+d)}\right]^2}{\frac{(a+b)(a+c)}{a+b+c+d}}$$

$$= \frac{\dfrac{(a^2 + ab + ac + ad - a^2 - ac - ab - bc)^2}{(a+b+c+d)^2}}{\dfrac{(a+b)(a+c)}{(a+b+c+d)}}$$

$$= \frac{(ad - bc)^2}{(a+b+c+d)(a+b)(a+c)}.$$

similarly, $\dfrac{[b - E(b)]^2}{E(b)} = \dfrac{(ad - bc)^2}{(a+b+c+d)(a+b)(b+d)}$,

$$\frac{[c - E(c)]^2}{E(c)} = \frac{(ad - bc)^2}{(a+b+c+d)(a+c)(c+d)}.$$

and $\dfrac{[d - E(d)]^2}{E(d)} = \dfrac{(ad - bc)^2}{(a+b+c+d)(b+d)(c+d)}.$

$$\therefore \chi^2 = \frac{(ad - bc)^2}{(a+b+c+d)}\left[\left\{\frac{1}{(a+b)(a+c)} + \frac{1}{(a+b)(b+d)}\right\} + \left\{\frac{1}{(a+c)(c+d)} + \frac{1}{(b+d)(c+d)}\right\}\right]$$

$$= \frac{(ad - bc)^2}{(a+b+c+d)}\left[\frac{b+d+a+c}{(a+b)(a+c)(b+d)} + \frac{b+d+a+c}{(a+c)(c+d)(b+d)}\right]$$

$$= (ad - bc)^2\left[\frac{1}{(a+b)(a+c)(b+d)} + \frac{1}{(a+c)(c+d)(b+d)}\right]$$

$$= \frac{(ad - bc)^2(a+b+c+d)}{(a+b)(a+c)(b+d)(c+d)} = \frac{(ad - bc)^2 n}{(a+b)(a+c)(b+d)(c+d)}.$$

Hence proved.

Yate's Correcrtion : We have already pointed out that the χ^2 distribution is a continuous distribution and χ^2 for testing goodness of fit and for testing independence of attribute is approximated to the χ^2-distribution when the expected cell frequencies are greater than 5. For values, less than 5, we use the method of pooling theoretical cell frequencies. But in case of 2 x 2 contingency table, the d. f. is 1 and the use of pooling method cannot be applied because it makes the d. f. zero which is meaningless. F. Yates (1934) provided a method of correction usually known as Yate's correction for continuity. This consist in adding 0.5 to the observed cell frequencies which are less than 5 and then adjusting for the remaining cell frequencies so that the marginal totals remain same.

For a 2 x 2 contingency table with cell frequencies $\begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array}$, the values of χ^2 after Yate's correction for continuity becomes

$$\chi^2 = \frac{n\left[\,|ad - bc| - \frac{n}{2}\right]^2}{(a+b)(a+c)(b+d)(c+d)} \qquad \text{........(**)}$$

Example 10.15 In an experiment with immunization of goats from anthrox the following results were obtained. Derive your inference on the efficiency of the vaccine.

	Died	Survived
Inoculated	2	10
Not Inoculated	6	6

Solution : After Yate's correction the contingency table becomes :

Table-10.3

	Died	Survived	
Inoculation	2.5	9.5	12
Not Inoculation	5.5	6.5	12
Total	8	16	24

We set up the hypothesis H_0 : The efficiency of vaccine over the disease is nil.

$$\chi^2 = \frac{24|6.5 \times 2.5 - 9.5 \times 5.5|^2}{12 \times 12 \times 8 \times 16} = \frac{24 \times 36^2}{12 \times 12 \times 8 \times 16} = 1.688 \text{ (app) with 1 d. f.}$$

The same result can be obtained by using (**).

The tabulated value of χ^2 with 1 d. f. at 5% level of significance is 3.81. It is seen that the calculated value of χ^2 with same d. f. is less than the tabulated value and hence it is insignificant and the hypothesis may be accepted.

(3.d) Test of significance of single variance

Let us suppose that we have a random sample of size n consisting of x_1, x_2,.........x_n drawn from a normal population. We want to test the null hypothesis that the population variance is σ^2_0, a specified value. i. e. $H_0 : \sigma^2 = \sigma_0^2$ (a specified value).

We know that the estimate of unknown population varaince σ^2 is,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \text{ where } \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The required test statistic is, $$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2} = \sum^{n} \frac{(x_i - \overline{x})^2}{\sigma_0^2} \qquad \text{..............(10.40)}$$

which is distributed as χ^2 with (n - 1) d. f.

The conclusion can be drawn as given the principles in (C).

Example 10.16 From a random sample of 21 values we calculate an estimate 4.5 for the variance of the population. Does this result support the hypothesis that the population variance is 10 ?

Solution : We set up the null hypothesis, $H_0 : \sigma^2 = 10$.

The test statistic is, $\chi^2 = \frac{20 \times 4.5}{10} = 9.00$, which is distributed as χ^2 with 20 d. f.

The tabulated value of χ^2 with 20. d. f. at 5% level of significance is 31.41, which is greater than the calculated value of χ^2 with 20 d. f. Hence the calculated value of χ^2 is insignificant and the hypothesis may be accepted.

(3.e) Test of significance of equality of several variances

Let us suppose that we have k independent samples each of size n_i (i = 1, 2,......., k) and they are randomly drawn from normal populations. We are to test the null hypothesis, $H_0 : \sigma_1^2 = \sigma_2^2 = \text{..........} = \sigma_k^2$.

Let s_i^2 (i = 1, 2,......., k) be the ith sample variance based on (n_i - 1) degrees of freedom and also let us define ;

$$s^2 = \frac{\sum_{i=1}^{k} (n_i - 1)s_i^2}{\sum_{i=1}^{k} (n_i - 1)} = \frac{\sum_{i=1}^{k} v_i s_i^2}{v}, \text{ where } v_i = n_i - 1 \text{ and } v = \sum_{i=1}^{k} v_i.$$

The required test statistic is,

$$\chi^2 = \frac{1}{M}\left\{ v \log_{10} s^2 - \sum_{i=1}^{k} v_i \log_{10} s_i^2 \right\} \qquad \text{........(10.41)}$$

which is approximately distributed as χ^2 with (k - 1).

The value of M is given by, $M = 0.43429 \left[1 + \frac{1}{3(k-1)} \left\{ \sum_{i=1}^{k} \frac{1}{v_i} - \frac{1}{v} \right\} \right]$.

This statistic is due to Bartlett.

The conclusion can be drawn as given the principles in (C).

Example 10.17 The estimated variances obtained from five independent samples and the corresponding degrees of freedom are given in Table-10.4.

Table-10.4

	Samples				
	1	2	3	4	5
s_i^2	2.50	3.20	5.61	4.34	5.83
v_i	7	6	3	4	8
$\log_{10} s_i^2$	0.39794	0.50515	0.74896	0.63749	0.76567

Test the null hypothesis, $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$,

Solution : Here, $s^2 = \frac{\Sigma v_i s_i^2}{v} = \frac{117.53}{28} = 4.20$. (app)

$\log_{10} s^2 = 0.62325$. $v \log_{10} s^2 = 17.451$.

$\sum_{i=1}^{5} v_i \log s_i^2 = 16.789$. $\qquad \sum_{i=1}^{5} \frac{1}{vi} = 1.01786$; $\quad \frac{1}{v} = 0.03571$.

Now, $M = 0.43429[1 + \frac{1}{12}(1.01786 - 0.03571)] = 0.43429 \left[1 + \frac{0.98215}{12} \right]$

$= 0.43429 \times 1.08185 = 0.46985.$

$$\therefore \chi^2 = \frac{1}{M}\left(v \log_{10} s^2 - \sum_{i=1}^{s} v_i \log_{10} s_i^2\right)$$

$$= \frac{1}{0.46985} \times 0.712 = 1.51 \text{ (app) with 4. d. f.}$$

The tabulated value of χ^2 with 4 d. f. at 5% level of significance is 9.488. Here the calculated value of χ^2 with same d. f. is seen to be insignificant and therefore, the hypothesis may be accepted.

(3.t) Test of significance of equality of several population proportions

Let us suppose that we have k groups of observations and the proportion for each group for possessing certain attribute A is obtained from k independent bionomial populations. We are to test the hypothesis that population proportions are same i. e. $H_0 : \pi_1 = \pi_2 = ... = \pi_k$:

where π_i is the ith populations proportions.
The sample from binomial populations may be arranged in Table -10.5.

Table-10.5

	No. of obs. possessing attribute A	Not A	Total
	r_1	$n_1 - r_1$	n_1
	r_2	$n_2 - r_2$	n_2
	|	|	|
	|	|	|
	r_k	$n_k - r_k$	n_k
Total	R	N - R	N

Let us calculate $P = \frac{R}{N}$; the required test statistic is

$$\chi^2 = \frac{1}{P(P-1)}\left\{\sum_{i=1}^{k} \frac{r_i^2}{n_i} - \frac{R^2}{N}\right\} \qquad ....(10.42)$$

which is approximately distributed as χ^2 with (k - 1) d. f.

The conclusion can be drawn as given the principles in (C).

Example 10.18 Five samples of seeds, selected at random one each from five lots were sown and their germination rates were observed. The results are given in Table - 10.6.

Table-10.6

	Samples					
	1	2	3	4	5	Total
Germinated	40	110	70	120	180	520
Not Germinated	10	40	30	30	20	130
Total	50	150	100	150	200	650

Test the equality of the proportions in the populations.

Solution : We set up the null hypothesis, $H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5$.

Here, $P = \frac{520}{650} = 0.8$, $1 - P = 0.2$ $\therefore P(1 - P) = 0.16$.

$$\sum \frac{r_i^2}{n_i} - \frac{R^2}{N} = \frac{40^2}{50} + \frac{110^2}{150} + \ldots\ldots + \frac{180^2}{200} - \frac{520^2}{650}$$

$= 419.7 - 416.0 = 3.7.$ (app)

Therefore, $\chi^2 = \frac{3.7}{0.16} = 23.1$ (app) with 4 d. f.

The tabulated value of χ^2 with 4 d. f. at 1% level of significance is 13.28. The calculated value of χ^2 is higly significant and hence the hypothesis may be rejected.

(3.g) Test of significance of equality of several correlation co-efficients

Let us suppose that $r_1, r_2, \ldots\ldots\ldots, r_k$ be the sample correlation co-efficients calculated from k independent random samples of sizes $n_1, n_2, \ldots\ldots\ldots, n_k$ respectively from seperate bivariate normal populations. We are to test the hypothesis that the populations correlation co-efficient are same i.e. $H_0 ; \rho_1 = \rho_2 = \ldots\ldots\ldots = \rho_k$.

We can obtain the value of $z_1, z_2, \ldots\ldots\ldots, z_k$ from Table No. 14, Page 139, Vide Biometrika Tables for Statisticians edited by E. S. Pearson and O. H. Hartley. Fisher's z transformation is given by,

$$z_i = \frac{1}{2} \log_e \frac{1 + r_i}{1 - r_i} ; i = 1, 2, \ldots\ldots\ldots, k.$$

These z_i's are normally distributed about a common mean

$$m = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \text{ and variance} = \frac{1}{n_i - 3}$$

The estimate of m is $\bar{z}$ which can be calculated by, $\bar{z} = \dfrac{\sum^{k} (n_i - 3) z_i}{\sum^{k} (n_i - 3)}$.

So that, $\dfrac{z_i - \bar{z}}{\frac{1}{\sqrt{(n_i - 3)}}} = (z_i - \bar{z})\sqrt{(n_i - 3)}$; $(i = 1, 2, \ldots\ldots, k)$ are

independent standardised normal variates with mean zero and variance 1.

Hence, $$\chi^2 = \sum_{i=1}^{k} (z_i - \bar{z})^2 (n_i - 3) \qquad \ldots\ldots\ldots\ldots(10.43)$$

which is distributed as χ^2 with (k - 1) d. f. This statistic is obtained by the additive property of χ^2-distribution. 1 d. f. is lost due to the estimate of m by $\bar{z}$.

Conclusion can be made as given the principles in (C).

Example 10.19 The correlation co-efficients between certain diet and rate of growing of fishes of numbers 10, 14, 16, 20, 25 and 28 from six independent ponds were found to be 0.318, 0.106, 0.253 0.340. 0.116 and 0.112. Test the homogenity of the population correlation co-efficients.

Solution : We set up the null hypothesis, $H_0 : \rho_1 = \rho_2 = \rho_3 = \rho_4 = \rho_5 = \rho_6$.

From z transformation we have the values of z_i's as

$z_1 = 0.3294$, $\quad z_2 = 0.1063$, $\quad z_3 = 0.2586$,
$z_4 = 0.3541$, $\quad z_5 = 0.1165$. $\quad z_6 = 0.1125$.

$$\therefore \bar{z} = \frac{\sum^{6}(n_i - 3)z_i}{\sum(n_i - 3)} = 0.1919. \text{ (app)}$$

Now, $\chi^2 = \sum^{6} (n_i - 3)(z_i - \bar{z})^2 = 0.1008$. (app) with 5. d. f.

The tabulated value of χ^2 with 5. d. f. at 5% level of significance is 11.070. Our calculated value is 0.1008, Hence the calculated value of χ^2 with same d. f. is insignificant and the hypothesis may be accepted.

4) F test : This test, given by Fisher and Snedecor, comes from the definition of F-distribution which reduces to s_1^2/s_2^2 with $(n_1 - 1)$ and $(n_2 - 1)$ d. f. where s_1^2 and s_2^2 denote two estimates of population variance σ^2, obtained from two independent random samples of sizes n_1 and n_2 respectively. Thus briefly, the statistic $F=s_1^2/s_2^2$ wich is distributed as F-distribution with $v_1 = (n_1 - 1)$ and $v_2 = (n_2 - 1)$ d. f. In the above test, greater of the two variances s_1^2 and s_2^2 is to be taken in the numerator and v_1 corresponds to the greater variance.

Uses : This test statistic is used mainly to test the null hypothesis regarding the equality of two population variances, homogeniety of independent estimates of population means, significance of sample correlation ratio and also for testing the linearity of regression.

The theoretical or tabulated value of F with different d. f. as well as different level of significance are given in Table No. V, Page - 53 and 55 Vide Statistical Tables for Biological Agricultural and Medical Research.

The conclusion can be drawn as below :

i) If the calculated value of F with v_1 and v_2 d. f. is smaller than the tabulated value of F with same d. f. at 5% level of significance then the calculated value of F is insignificant and the null hypothesis may be accepted. (D)

ii) If the calculated value of F with v_1 and v_2 d. f. is greater than the tabulated value of F with same d. f. at 5% level of significance but smaller than the tabulated value of F with same d. f. at 1% level of significance, then the calculated value of F is significant and the hypothesis may be rejected at 5% level of significance.

iii) If the calculated value of F with v_1 and v_2 d. f. is greater than the tabulated value of F with same d. f. at 1% level of significance then the calculated value of F is highly significant and the null hypothesis may be rejected.

N. B. : Significant value of any test statistic (calculated) is denoted by* and the highly significant value of the same is denoted by**.

Applications of F test

(4.a) Test of significance for equality of two population variances

Let us suppose that $x_1, x_2, \ldots\ldots, x_{n_1}$ and $y_1, y_2, \ldots\ldots, y_{n_2}$ be two independent random samples of size n_1 and n_2 drawn from two normal populations. We have to test the null hypothesis that the two population variances are same i.e. $H_0 : \sigma_1^2 = \sigma_2^2$.

The estimates of the population variance are

$$s_x^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \overline{x})^2 \text{ and } s_y^2 = \frac{1}{n_2 - 1} \sum_{1}^{n_2} (y_j - \overline{y})^2$$

where $\overline{x} = \dfrac{\sum^{n_1} x_i}{n_1}$ and $\overline{y} = \dfrac{\sum^{n_2} y_j}{n_2}$.

The required test statistic is $F = \dfrac{s_x^2}{s_y^2}$ ....(10.44)

which is distributed as F with $v_1 = (n_1 - 1)$ and $v_2 = (n_2 - 1)$ d. f. In the above test, we consider $s_x^2 > s_y^2$.

The conclusion can be made as given the principles in (D).

Exmple 11.20 Two random samples drawn from two normal populations are :

Sample 1 : 20, 16, 26, 27, 23, 22, 18, 24, 25, 19.

Sample 2 : 27, 33, 42, 35, 32, 34, 38, 28, 41, 43, 30, 37.

Obtain estimates of the variances of the population and test whether the two populations have the same variance.

Solution : We set up $H_0 : \sigma_1^2 = \sigma_2^2$.

Mean of Sample 1, $\overline{x}_1 = \dfrac{\sum x_{1i}}{n_1} = \dfrac{220}{10} = 22.$

Variance of sample 1, $s_1^2 = \dfrac{\sum^{n_1} (x_{1i} - \overline{x}_1)^2}{n_1 - 1} = \dfrac{120}{9} = 13.33.$

Mean of Sample 2, $\overline{x}_2 = \dfrac{\sum^{n_2} x_{2j}}{n_2} = \dfrac{420}{12} = 35.$

Variance of sample 2, $s_2^2 = \dfrac{\sum^{n_2} (x_{2j} - \overline{x}_2)^2}{n_2 - 1} = \dfrac{314}{11} = 28.55.$

The required test statistic is, $F = \dfrac{s_2^2}{s_1^2} = 2.14$ (app). Since, $s_2^2 > s_1^2$.

The tabulated value of F with (11, 9) d. f. at 5% level of significance is 3.1. Our calculated value is 2.14. Hence the calculated value of F is insignificant and the hypothesis may be accepted.

(4.b) Test of significance for homogenity of population means

Let us suppose that we have k(k>2) independent random samples drawn from normal populations. We want to test the null hypothesis,

$H_0 : \mu_1 = \mu_2 = ... = \mu_k$ where μ_i is the mean of the ith population (i = 1, 2,.., k).

The sample observations are arranged as below :

	1st Sample	2nd Sample	....	kth Sample
	x_{11}	x_{21}		x_{k1}
	x_{12}	x_{22}		x_{k2}
	|	|		|
	x_{1n_1}	x_{2n_2}		x_{kn_k}
Total	T_1	T_2		T_k
Mean	$\overline{x_1}$	$\overline{x_2}$		$\overline{x}_k$

Let $T = \sum\limits_{i=1}^{k} T_i = \sum\limits_{i=1}^{k} \sum\limits_{j=1}^{ni} x_{ij}$, $N = \sum\limits_{i=1}^{k} n_i$ and $\overline{x} = \frac{T}{N}$

In the above samples, the total sum of squares, the total sum of squares (S_t) can be partitioned into two componenets namely between sum of squares (S_b) and within sum of squares (S_w).

The test statistic is, $F = \frac{S_b/(k-1)}{S_w/(N-k)}$ ....(10.45)

which is distributed as F-distribution with (k - 1) and (N - k) d. f.

The usual method of calculation of different components of sum of squares are as follows :

$S_t = \sum\limits_{i}\sum\limits_{j}(xij - \overline{x})^2 = \sum\limits_{i}\sum\limits_{j} x_{ij}^2 - \frac{T^2}{N}$ $\quad S_b = \sum n_i(\overline{x}_i - \overline{x})^2 = \sum\limits_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{N}$

$\therefore S_w = S_t - S_b = \sum\sum(x_{ij} - \overline{x}_i)^2.$

The conclusion can be drawn as given the principles in (D).

Remark : The above technique is usually called "Analysis of Variance" for one-way classification data. An elaborate discussion on it and also for more than one-way classification data is given in the next chapter.

Example 10.21 10 varieties of wheat are given in 3 plots each and following yields in kg. per plot are obtained. Test the homogenity of the population means of different varieties.

Table-10.7

Plot/Variety	1	2	3	4	5	6	7	8	9	10
1	7	7	14	11	9	6	9	8	12	9
2	8	9	13	10	9	7	13	13	11	12
3	7	6	16	11	12	5	12	11	11	11
Total	22	22	43	32	30	18	34	32	34	31

Solution : We set up $H_0 : \mu_1 = \mu_2 = ... = \mu_{10}$ where μ_i indicates mean yield of ith variety, we calculate, $\Sigma T_i = 298 = T$. and $N = 30$

Total S. S. $(S_t) = \Sigma\Sigma x_{ij}^2 - \frac{T^2}{N} = 203.87$

Between variety S.S. $(S_b) = \frac{\Sigma T_i^2}{3} - \frac{T^2}{N}$

$= \frac{1}{3}(22^2 + 22^2 + ...+ 31^2) - \frac{298^2}{30} = 160.54.$

Within variety S.S. $(S_w) = 203.87 - 160.54 = 43.33$

The test statistic is, $F = \frac{160.54/9}{43.33/20} = 8.22$ (app) with (9,20) d. f.

The calculated value of F with (9, 20) d. f. is greater than the tabulated value of F with same d. f. at 1% level of significance. Hence the calculated value of F is highly significant and the hypothesis may be rejected.

(4.c) Test of significance of an observed correlation ratio

Let us suppose that we have a random sample of size N from a bivariate normal population. The observations are arranged in h arrays. We are to test the null hyppothesis that the population correlation ratio is zero, i. e. $H_0 : \eta = 0$.

The required test statistic is $F = \frac{\eta^2}{1-\eta^2} \times \frac{N-h}{h-1}$ ....(10.46)

which is distributed as F with (h - 1), (N - h) d. f.

The conclusion can be drawn as given the principles in (D).

Example 10.22 A random sample of 80 pairs of values from a bivariate normal population grouped in 10 arrays of y's gives a correlation ratio $\eta_{yx} = 0.2$. Is it significant of association between the variates?

Solution : We set up the null hypothesis that the population correlation ratio is zero i. e. $H_0 : \eta = 0$.

Here $N = 80$, $h = 10$, $\eta_{yx} = 0.2$.

The test statistic is $F = \frac{0.04}{1-0.04} \times \frac{70}{9} = \frac{2.80}{8.64} = 0.32$ (app) with (9, 70) d. f.

Since the calculated value of F with (9,70) d.f. is smaller than the tabulated value of F with same d.f. at 5% level of significance, the calculated value is insignificant and therefore, the hypothesis may be accepted.

(4.d). Test of significance of linearity of regression

Let us suppose that we have a random sample of size N arranged in h arrays, taken from a bivariate normal population. We are to test the null hypothesis of linearity of regression.

The required test statistic is $F = \frac{\eta^2 - r^2}{1-\eta^2} \times \frac{N-h}{h-2}$ ....(10.47)

which is distributed as F distribution with (h - 2), (N - h) d. f.

Here η is the correlation ratio and r is the correlation co-efficient.

Example 10.23 A random sample of 100 pair from a bivariate normal population when grouped in 10 array of y's gives $r = 0.4$ and $\eta = 0.5$. Are these results consistent with the assumption of linearity of regression?

Solution : We set up the null hypothesis that the regression is linear.

Here $N = 100$, $h = 10$, $r = 0.4$, $\eta = 0.5$

The test statistic is, $F = \frac{0.25 - 0.16}{1 - .25} \times \frac{90}{8}$

$$F = \frac{0.09 \times 11.25}{0.75} = \frac{1.0125}{0.75} = 1.35, \text{ with } (8, 90) \text{ d. f.}$$

The calculated value of F with (8, 90) d.f. is less than the tabulated value of F at 5% level of significance. Hence the calculated value is insignificant and the hypothesis may be accepted.

(4.c) Test of significance of an observed multiple correlation co-efficient

Let us suppose that R be the multiple correlation co-efficient of order k in random sample of size N from a (k + 1) variate population. We are to test the null hypothesis that the population multiple correlation co-efficient is zero i. e. $H_0 : R = 0$.

The required test statistic is, $F = \frac{R^2}{1 - R^2} \cdot \frac{N - k - 1}{k}$ ....(10.18)

which is distributed as F with k, (N - k - 1) d. f.

The conclusion can be made as given the principles in (D).

Example 10.24 For a sample of 30 sets of values from a normal population, $R_{2.31}$ is found to be 0.5. Test that the population multiple correlation co-efficient is zero.

Solution : We set up the null hypothesis that the population multiple correlation co-efficient is zero.

The test statistic is, $F = \frac{0.25}{1 - 0.25} \times \frac{30 - 2 - 1}{2} = \frac{0.25}{0.75} \times \frac{27}{2} = 4.5$ with (2,27) d. f.

The calculated value of F with (2, 27) d. f. is seen to be greater than the tabulated value of F with same d. f. at 5% level of significance. Hence the calculated value of F is significant and the hypothesis may be rejected at 5% level of significance.

11. DESIGN OF EXPERIMENTS

11.1 Introduction :

By the word experiment we mean a process to have a series of trials or observations taken under some condition specified by the experimenter to confirm or disprove something doubtful and also to discover some unknown principles or effects or to test, establish or illustrate some suggested known truth. The design of experiments mean the logical construction of experiment to select the pattern of collecting data to suit the above purposes.

Broadly experiment can be divided into two parts, absolute and comparative. In absolute experiment, the characteristic is fixed and observations are collected to make the best estimate of that. Design of sample survey is an example of absolute experiment. On the other hand, comparative experiments are designed to compare the effects of two or more objects on some population characteristics. Thus design of experiments refer to comparative experiments.

Before going in detail of this chapter we are giving below the explanations of the terms used in different places.

Treatments : Different procedures under comparison in an experiment may be termed as treatments. For example, in agricultural experiments different varieties of a crop, different levels of fertilizer may be considered as treatments. In medical experiment different doses of a medicine or diets are the treatments.

Experimental Unit : It is the experimental meterial to which we apply the treatments and on which we make observation on the variable under study is termed as experimental unit. A plot of land and a batch of seeds are experimental units in agricultural experiments whereas patients in a hospital or a group of pigs may be considered as experimental unit in medical experiments.

Blocks : In most of the times we divide the whole experimental unit into homogeneous sub-groups or strata which as a whole may be termed as blocks. A number of homogeneous plots in a strip constitute a block in an agricultural experiment where as the patients of same symptoms having same age-group, same sex etc. may constitute a block in a medical experiment.

Yields : The measurements of the variable under study on different experimental plots are termed as yields.

Experimental Error : The yields of an experiment are usually influenced by some extraneous variations may or may not be controlled by the experimenter. The uncontrolled variations are often called the experimental errors. For a homogeneous experimental unit devided into different plots of equal sizes and different treatments are applied to these plots ; the yields of these plots wil not be same. The difference of the yields may be due to difference of treatments or due to difference of inherent soil structure or fertility condition of the soil. In field experiment, experience tells us that even same treatments are used on all the plots, the yield would still vary due to these sources of variations. Such variations from plot to plot are due to random compound and beyond human control, isrefered to experimental error.

The error includes all types of extraneous variations which are due to the following factors :

i) inherent variability in the experimental material to which the treatments are applied.
ii) the lack of uniformity in the methodology of conducting experiment.
iii) lack of representativeness of the sample to the population under study.

Replication : The repeated application of treatment under investigation is known as replication. Detail explanation and uses of replication is given in the principles of experimental design.

Precision : The reciprocal of the variance of the treatment mean is termed as precision or the amount of information in the design. In an experiment, if a treatment is replicated r times, then the precision is given by $\frac{r}{\sigma^2}$ where σ^2 is the error variance per unit.

Efficiency of a Design : Let D_1 and D_2 be two designs with error variances per unit σ_1^2 and σ_2^2 and replications r_1 and r_2 respectively. The variances of the differences between two treatment means are given by $\frac{2\sigma_1^2}{r_1}$ and $\frac{2\sigma_2^2}{r_2}$ for D_1 and D_2 respectively. We define the ratio of the informations, $E = \frac{r_1}{2\sigma_1^2} \div \frac{r_2}{2\sigma_2^2}$ as the efficiency of the design D_1 in comparison to D_2, if $E = 1$; D_1 and

D_2 are equally efficient, if, $E > 1$ ($E < 1$) D_1 is said to be more (less) efficient than D_2.

Contrast : Let $y_1, y_2, \ldots, y_n$ be the n observations, then the linear function $c = l_1y_1 + l_2y_2 + \ldots + l_ny_n$ is a contrast of y_i's if l_i's are some numbers such that $\sum_{i=1}^{n} l_i = 0$. The sum of squares of the contrast c is defined by $\frac{c^2}{\sum l_i^2}$.

Orthogonal Contrasts : Two contrasts $c_1 = \sum_i l_iy_i$ and $c_2 = \sum m_iy_i$ are said to be orthogonal if $\sum_i l_im_i = 0$. When there are more than two contrasts they are said to be mutually orthogonal, if they are orthogonal pair wise.

Important steps in Design of Experiments : Following are the important steps to be considered by an experimenter to have a good design of experiment.

The statement of the problem should be clearly defined. In that case, he can understand what to do and how to tackle the problem.

Formulation of the hypothesis should be done properly and thus the method of collection of data can be determined. For these two steps we can think of any previous experience whose reference can be made to throw some light and adequate information for possible results from the point of view of statistical theory on future experiment may be required.

The experiment should be conducted accordingly and proper statistical techniques are to be applied on the data.

Drawing of valid conclusions is the crutial part of design of experiment, so careful considerations are to be given for the validity of the conclusions for the population of objects or events to which they are to apply. Also evaluation of the whole investigation and comparison of the results can be done with similar past investigation.

Principles of Design of Experiment : According to Prof. R. A. Fisher, the basic principles of design of experiments are (a) randomisation, (b) replication and (c) error control. The explanations of the terms are given below :

(a) Randomisation : At first the treatments and experimental plots of the experiment are decided. Randomisation means that for an objective comparison it is necessary that the treatments be alloted randomly to different experimental plots to avoid any type of personal or subjective error i. e. without giving higher importance to any of the treatments. It also ensures independence of the observations which is necessary for drawing valid inference by applying statistical techniques.

There are numbers of ways of randomisation depending on the nature of the design of experiment. The individual process of randomisation will be described in appropriate cases.

(b) Replication : The repetition of the treatments under investigation to more than one experimental plots is known as replication. For example, a treatment is alloted to 'r' plots of an experimental unit then it can be said that the treatment is replicated 'r' times. Replication is necessary to increase the accuracy of the estimates of the treatment effects, it also provides an estimate of error variance. It is seen that the precision increases if the replication increases, but it cannot be increased indefinitely due to limited resources i. e. time, money, skilled personnels etc. The number of replications, therefore, depend on the expenditure and the degrees of precision. Sensitivity of statistical methods for drawing inferences also depend on the number of replications.

Determination of Number of Replication : If $\overline{y_1}$ andn $\overline{y_2}$ be the mean effects of two treatments replicated r_1 and r_2 times respectively, then

var ($\overline{y_1} - \overline{y_2}$) = var ($\overline{y_1}$) + var ($\overline{y_2}$), since the co-variance term vanishes due to independence of observations.

$\therefore$ Var ($\overline{y_1} - \overline{y}_2$) $= \frac{\sigma^2}{r_1} + \frac{\sigma^2}{r_2} = \frac{2\sigma^2}{r}$ if $r_1 = r_2 = r$ and σ^2 is the usual error variance. Therefore, the standard error of ($\overline{y_1} - \overline{y_2}$) is equal to $\sigma\sqrt{\frac{2}{r}}$

For testing the equality of two means for large sample under the usual assumption $\frac{\overline{y_1} - \overline{y_2}}{\text{St. Er} (\overline{y_1} - \overline{y_2})}$ is a (N(0, 1) variate.

For small sample the estimate of σ^2 is done and the test statistic is distributed as t with d.f. depending on the divisor of the estimate of σ^2. i. e. s^2. Therefore, for a certain level of significance, say at α% and with d. f., the critical value of t_α can be obtained from the t -table.

Then, $t_\alpha = \dfrac{|d|}{s\sqrt{\frac{2}{r}}}$ or, $r = \dfrac{2t^2_\alpha s^2}{d^2}$, where $|d| = \overline{y_1} - \overline{y_2}$.

Thus the number of replications, r is obtained.

(c) Error Control : Though every experiment would provide an estimate of error variance, it is not desirable to have a large experimental error. The measure for reducing the error variance are usually called error control or local control. One such measure is to make experimental units homogeneous, another method is to form experimental units into several homogeneous groups usually called blocks, allowing variation among the groups. Different methods of forming groups of homogeneous plots for allotment of groups of treatments are used now a days for the estimation of treatment effect precisely. In short, the aim of error control is to reduce the error by modifying the allocation of treatments to the experimental units.

Models and Analysis of Variance : A statistical model is generally a linear relation of the effects of a member of factors with different levels in an experiment and also one or more terms representing error effects. The effects of any factor may be random or fixed depending on the method of selecting the levels of the factors. For example, if there are number of variations of a crop of which one variety is selected at random then the varietal effect would be random, while the effect of two well defined levels of irrigation are fixed as each irrigation level can be reasonably taken to have a fixed effect.

The models of experiments are of three types namely (i) fixed effect model (ii) random effect model and (iii) mixed effect model.

A model in which each of the factors has fixed effect and only the error effect is random, is called fixed effects model. The random effect model is that one, in which all the effect in a model are random. The model in which some factors have fixed effects and some factors have random effects, is called mixed effect model.

In this text, we shall consider only the fixed effect models whose main objectives are to estimate the effects, to obtain a measure of variability among the effects of each of the factors and finally to find the variability among the error effects.

The data of usual design of experiment can be classified as follows :

When a set a observations is distributed over the different levels of a factor, they form one-way classified data. Let us consider one factor at k levels. Let there be n_i observations denoted by y_{ij} (i = 1, 2,.......k, j = 1, 2,......n_i) against the ith level. Then the observations y_{ij} classified in k levels of the factor are said to form one-way classified data. Similarly, if we take two factors simultaniously, say, A and B at number of levels k and r respectively, then there are (k x r) cells, each of which is defined by one level of A and one level of B. Let there be n_{ij} observations in the (i, t)th cell defined by the ith level of A and jth level of B. Let y_{ijl} denote the lth observation in the (i, j)th cell. Then the data y_{ijl} (i = 1, 2,.........k, j = 1, 2,.......r and l = 1, 2,.......n_{ij}) arranged in the (k x r) groups are called two-way classified data.Similarly, in general, m-way classified data can be defined by using levels of m-factors simultaniously.

Now considering two factors A and B involving in an experiment without intraction, the fixed effect model for two-way classified data can be written as, $y_{ij} = \mu + a_i + b_j + e_{ij}$; where y_{ij} is the observation coming from ith and jth levels of two factors respectively involved in the experiment, a_i is the effect of the ith level of factor A, b_j is the effect of the jth level of the factor B and e_{ij} is the error component which is assumed to be independently and normally distributed with zero mean and a constant variance σ^2. These assumptions regarding the behaviour of e_{ij} are necessary for appropriate statistical methodology for drawing valid inference. The adopted methodology is the analysis of variance technique by which inference is drawn by applying F-test.

One further assumption is the additivity of the effects in the model. This assumption is generally satisfied except for some less known situation. For that Tukey's test for additivity is available.

The models may be of different types depending on the nature of the data i.e. the number of factors involved in the experiment. The above model is appropriate for two-way classified data without any interaction among the effects of the factors. For m different factors we can have m-way classified data and accordingly the models can be written.

The analysis of variance is the systematic procedure of partitioning the total variation present in a set of observation, into number of components associated with the nature of classification of data. For one-way classified data the total variation can be partitioned into two components namely variation due to the single factor and the other is due to error variation. This error includes all possible extraneous error components. For two-way classified data involving factors A and B the total variation can be partitioned into three components e.g. variation due to A, variation due to B and error variation. Similarly for three-way classified data involving three factors A, B and C, the total variation can be partitioned into four components e. g. variation due to A, variation due to B, variation due to C and error variation. The techniques of spliting of total variations are given in appropriate places. The spliting helps to get mean square due to different components and thus the relevant tests can be performed. For detail discussions of different types of analysis of variance Das and Giri (1979) can be referred.

11.2 Basic Designs

Basic designs include completely randomised design (C. R. D), randomised block design (R. B. D), and Latin square design (L. S. D). Each of these designs is described one after another with relevant extensions.

Completely Randomised Design (C.R.D) : It is the simplest design where only two principles viz, replication and randomisation are used in field experiment. In this design, the whole experimental material should be homogeneous in nature and is divided into number of experimental plots depending on the number of treatments and the number of replications for each treatment.

The design is useful mainly for laboratory or green house experiments whereas its uses in field experiment is limited. Complete flexibility is allowed in this design i. e. any number of treatments may be replicated any number of times. Missing plot and unequal replicates donot create any difficulty in analysing the data in this design. The principal objection to the use of this design is on the ground of accuracy when the plots are considered to be homogeneous wrongly.

Lay-out : The lay-out of a design indicates the placement of treatments to the experimental plots according to the condition of the design.

Let us consider an example to illustrate the layout of a C. R. D with 3 treatments A, B and C replicated 5, 3 and 2 times respectively. Here the experimental unit is to be devided into 10 equal plots and they are to be numbered. From Random Number Tables ten 3 degits numbers are taken and ranked. We take additional numbers in case of ties. From the ranked numbers first 5 numbered plots are considered to allote treatment A. Similarly treatments B and C can be alloted and thus the lay-out of C. R. D is obtained. For equally replicated treatments, similar method of randomisation can be carried out.

Analysis : The additive model for completely ranc ›mised design with unequal obstractions is

$y_{ij} = \mu + t_i + e_{ij}$; $(i = 1, 2, ...k ; j = 1, 2, ...n_i)$

where y_{ij} is the observations of the ith treatment in the jth replicate,

μ = general mean,

t_i = effect due to ith treatment,

e_{ij} = random error components which are assumed to be normally, independently distributed with 0 mean and variance σ^2.

Let there be k treatments and the ith treatment be replicated n_i times. Let $y_{i.}$ be the total of the observations corresponding to ith treatment and $y_{..}$ be the grand total of all the observations i e.

$y_{i.} = \sum_j y_{ij}$; $y_{..} = \sum_i y_{i.} = \sum_i \sum_j y_{ij}$ and total number of observations, $N = \sum_i n_i$.

The least square estimate of μ and t_i can be obtained by minimising the error sum of squares, denoted by $\sum_i \sum_j e_{ij}^2 = S = \sum_i \sum_j (y_{ij} - \mu - t_i)^2$

The normal equations are, $\sum\sum y_{ij} = N\mu + \sum n_i t_i$ and

$$\sum_j y_{ij} = n_i \mu + n_i t_i$$

Out of these two equations only one is independent because taking summations over i in the second equation we get the first one. To have unique solution we have to impose restriction $\sum n_i t_i = 0$

Now, we have the solutions as follows :

$\hat{\mu} = \sum\sum y_{ij}/N = \overline{y}_{..}$ where $\overline{y}_{..}$ is the grand mean of all the observations.

and $\hat{t_i} = \overline{y}_{i.} - \overline{y}_{..}$ where $\overline{y}_{i.}$ is the mean of the obstrations corresponding to ith treatment.

To show that the estimates are independent, we have,

$$Cov(\hat{\mu}, \hat{t}_i) = Cov\{\overline{y}_{..}, (\overline{y}_{i.} - \overline{y}_{..})\}$$

$$= Cov(\overline{y}_{i.}, \overline{y}_{..}) - var(\overline{y}_{..}) = \frac{n_i\sigma^2}{N n_i} - \frac{\sigma^2}{N} = 0,$$

showing that the estimates are independent.

The total sum of squares, in this case, can be partitioned into two components as follows :

$$\sum_i\sum_j(y_{ij} - \overline{y}_{..})^2 = \sum_i\sum_j\{(y_{ij} - \overline{y}_{i.}) + (\overline{y}_{i.} - \overline{y}_{..})\}^2$$

$$= \sum_i\sum_j(y_{ij} - \overline{y}_{i.})^2 + \sum n_{i.}(\overline{y}_{i.} - \overline{y}_{..})^2, \text{ the product term vanishes.}$$

Thus we get, Total S.S. = Within S.S. + Between S.S. Within S.S. and Between S.S. are usually called Error S.S. and treatment S.S. respectively.

Now, we are to show that different components of sum of squares follow χ^2-distribution with appropriate degrees of freedom.

We know, $y_{ij} = \mu + t_i + e_{ij}$

$$\overline{y}_{i.} = \mu + t_i + \overline{e}_{i.}$$

$$\overline{y}_{..} = \mu + \overline{t} + \overline{e}_{..}$$

Now, Treatment S. S. $= \sum_i n_i(\overline{y}_{i.} - \overline{y}_{..})^2$

$$= \sum_i n_i(\mu + t_i + \overline{e}_{i.} - \mu - \overline{t} - \overline{e}_{..})^2$$

$$= \sum n_i(t_i - \overline{t} + \overline{e}_{i.} - \overline{e}_{..})^2$$

$$= \sum n_i(t'_i + \overline{e}_{i.} - \overline{e}_{..})^2 \text{ ; considering } t_i - \overline{t} = t'_i$$

$$\sum n_i(t'^2_i + \overline{e}_{i.}^2 + \overline{e}_{..}^2 - 2t'_i\overline{e}_{..} + 2t'_i\overline{e}_{i.} - 2\,\overline{e}_{i.}\,\overline{e}_{..})$$

Taking expectation on both the sides and assuming $t'_i = 0$ under null hypothesis, $H_0 : t_1 = t_2 = \ldots\ldots = t_k$ we have,

$$E\left[\sum_i n_i(\overline{y_i}. - \overline{y}..)^2\right] = E\sum n_i \overline{e_i}.^2 + E\sum n_i \overline{e}..^2 - 2E\sum n_i \overline{e_i}.\ \overline{e}..$$

$$= \sum n_i \frac{\sigma^2}{n_i} + N\frac{\sigma^2}{N} - 2Nk\frac{\sigma^2}{Nk} = k\sigma^2 + \sigma^2 - 2\sigma^2 = \sigma^2(k-1).$$

or, $E\sum \dfrac{(\overline{y_i}. - \overline{y}..)^2}{\sigma^2/n_i} = k - 1$

which implies that $\sum \dfrac{(\overline{y_i}. - \overline{y}..)^2}{\sigma^2/n_i}$ is distributed as χ^2 with (k - 1) d. f.

Similarly Error S.S. $= \sum_i\sum_j (y_{ij} - \overline{y_{i.}})^2 = \sum_i\sum_j(\mu + ti + e_{ij} - \mu - t_i - \overline{e_i}.)^2.$

$$= \sum_i\sum_j (e_{ij} - \overline{e_i}.)^2$$

Proceeding as above and taking expectation on both the sides we have

$E\sum_i\sum_j \dfrac{(y_{ij} - \overline{y_{i.}})^2}{\sigma^2} = N - k$ which implies that $\sum\sum \dfrac{(y_{ij} - \overline{y_{i.}})^2}{\sigma^2}$ is distributed as

χ^2 with (N - k) d. f.

From the additive property of χ^2 it can be said that $\sum_i\sum_j \dfrac{(y_{ij} - \overline{y_{..}})^2}{\sigma^2}$ is also distributed as χ^2 with N - k + k - 1 = N - 1 d.f. It can be shown independently also.

Thus it is seen that each of the components of sum of squares is independently distributed as χ^2 with appropriate d. f.

Now, considering H_0, we have the test criterion

$$F = \frac{\sum n_i(\overline{y_i}. - \overline{y..})^2/(k-1)}{\sum_i\sum_j(y_{ij} - \overline{y_{i.}})^2/(N-k)} = \frac{\text{M. S. due to Treatment.}}{\text{M. S. due to Error}},$$

which is distributed as F with (k - 1) and (N - k) d. f.

Method of calculation of different sum of squares :

Total S.S. = $\sum_i\sum_j y_{ij}^2$ - C. F. = T_o, say, where C. F. = $\frac{y_{..}^2}{N}$

Treatment S.S. = $\sum_i \frac{y_{i.}^2}{n_i}$ - C.F. = T, say.

Error S.S. = Total S.S. - Treatment S.S. =T_0 - T = E, say.

Now the analysis of variance table can be furnished for testing the null hypothesis H_0 : Effect of all the treatments are same.

Table-11.1

ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F
Treatment	k - 1	$T = \sum_i \frac{y_i^2}{n_i}$ - C. F.	$T' = T/(k-1)$	T'/E'
Error	N - k	$T_o - T = E$	$E' = E/(N-k)$	
Total	N - 1	$T_o = \sum\sum y_{ij}^2$ - C. F.		

If the calculated value of F with (k - 1) and (N - k) d. f. is greater than the tabulated value of F with same d. f. and at 100α% level of significance, then the hypothesis may be rejected i.e. the effects of all the treatments are not same. Otherwise the hypothesis may be accepted.

Note : When the number of replications per treatment is same, say, n, then the normal equations become ;

$\sum_i\sum_j y_{ij} = N\mu + n \sum_i t_i$,

$\sum_j y_{ij} = n\mu + nt_i$, where we take N = nk. and the estimates are as usual.

The partitioning of the total sum of squares is

Total S.S. $= \sum_i\sum_j(y_{ij} - \overline{y_{..}})^2 = \sum_i\sum_j(y_{ij} - \overline{y_{i.}})^2 + n\sum_i(\overline{y_{i.}} - \overline{y_{..}})^2$

The calculations of the treatment sum of squares can be obtained by the following way .

Treatment S.S. = $\frac{1}{n}\sum y_{i.}^2$ - C. F

Example 11.1 A feeding trial with 3 feeds namely i) Pasture (control) ii) Pasture and Concentrates and iii) Pasture, Concentrate and Minerals was conducted to a certain variety of ewe lambs with same age, body weight and sex etc. 37 ewe lambs are selected for the purpose. The weight records of the total wool yields (in kg) of first two cliping were obtained. The purpose of the experiment is to serve whether the feeds have any effect on the wool yield.

Feed I : 50.5 53.6, 78.8, 65.4, 80.4, 95.3, 50.5, 52.5, 80.6, 75.2. 68.6, 69.7, 71.2, 73.1, 95.2.

Feed II : 63.9, 52.0, 78.8, 67.0, 80.4, 67.3, 53.6, 59.1, 63.5, 60.9.

Feed III : 59.1, 71.3, 69.1, 55.3, 61.9, 63.5, 76.1, 59.5, 62.3, 57.3, 61.5, 68.3.

Solution : We have,

	Total
Feed I :	106.6(15)
Feed II :	946.5(10)
Feed III :	765.2(12).
Grand Total :	2472.3(37)

The figure in the bracket indicates number of items.

Correction factor (C. F.) $= \frac{2472.3^2}{37} = 165196.41.$

Total S.S. $= 50.5^2 + 53.5^2 + \ldots\ldots + 68.5^2 -$ C. F.

$= 169756.47 -$ C. F. $= 4560.06.$

S.S. due to feed $= \frac{1060.6^2}{15} + \frac{946.5^2}{10} + \frac{765.2^2}{12} -$ C. F. $= 165581.97 -$ C. F. $= 385.56$

$\therefore$ Error S.S. = Total S.S. - S.S. due to feed. $= 4560.06 - 385.56 = 4174.5.$

We are to test null hypothesis H_0 : The effect of all the feeds are same.

Table-11.2

ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F	5%F
Treatment	2	385.56	192.78	1.57	3.284
Error	34	4174.5	122.77		
Total	36				

Since the calculated value of F is smaller than the tabulated value at 5% level of significance, the value is insignificant and the hypothesis may be accepted.

Randomised Block Design (R.B.D) : In many real situations it may not be possible to get homogeneous experimental unit as a whole but it is usually possible to get homogeneous groups of plots which are termed as blocks. By this way, we can control the variability in one more direction by assigning the treatments at random to each plot of the block, giving a design known as randomised block design. In this design the number of plots per block is the number of treatments and the number of blocks will determine the number of replications.

This is a popular design for its simplicity, flexibility and validity and can be applied with moderate number of treatments (<10). By means of grouping, the efficiency of the design can be increased than that of C.R.D. Any number of treatments and any number of replications can be carried out in this type of design but the number of replications for each treatment must be same. The statistical analysis is straight forward even if one or more observations are missing as given by Glenn and Kramer (1958) and Mitra (1959).

With the increase in number of treatments the block size increases and thus the homogeneity of block reduces resulting larger error components.

Lay-out : Let there be k treatments each replicated r times in the design. Therefore, the total number of plots required in this design is kr, which are arranged into r homogeneous groups called blocks each of size k. The number of plots per block is equal to the number of treatments and the number of replications are equal to the number of blocks determined by the available resources. All the blocks and the plots must be of same size. Randomisation of the treatments is done independently in each of these blocks.
Let us consider an example of randomisation of 5 treatments A, B, C, D and E in a single block. The treatments are numbered in any order, say A is assigned 1, B is assigned 2 and so on. From Random Number Table we take at least five 3 digits number and are ranked and their order is say, 3, 1, 4, 2 and 5. Now in the block 3rd treatment C is placed at the first plot, 1st treatment A is placed in the second plot and so on. Thus the randomisation in the block is completed. Seperate randomisation is done for each block.

Analysis : For analysis of data in this type of design the linear additive model be, $y_{ij} = \mu + t_i + b_j + e_{ij}$; $(i = 1, 2, \ldots\ldots k ; j = 1, 2, \ldots\ldots r)$

where y_{ij} is the observation for the ith treatment in the jth block.
μ is the general mean effect,
t_i is the effect due to ith treatment,
b_j is the effect due to jth block, and

e_{ij}, random error components which are assumed to be independently and normally distributed with zero mean and constant variance σ^2.

Let $y_{i\cdot} = \sum_j y_{ij}$; $y_{\cdot j} = \sum_i y_{ij}$; $y_{\cdot\cdot} = \sum_i y_{i\cdot} = \sum_j y_{\cdot j} = \sum_i \sum_j y_{ij}$. and

$\bar{y}_{i\cdot} = \frac{y_{i\cdot}}{r}$; $\bar{y}_{\cdot j} = \frac{y_{\cdot j}}{k}$ and $\bar{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{rk}$.

The least square estimate of t_i and b_j can be obtained by minimising the error sum of squares denoted by, $\sum_i \sum_j e_{ij}^2 = S = \sum_i \sum_j (y_{ij} - \mu - t_i - b_j)^2$.

In this case we get three normal equations which can be solved by imposing two restrictions, $\sum_i t_i = \sum_j b_j = 0$ giving the solutions as below :

$$\hat{\mu} = \bar{y}_{\cdot\cdot}\,;\ \hat{t}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \text{ and } \hat{b}_j = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$$

To show that the estimates are independent we have,

$$\text{cov}\,(\hat{\mu}, \hat{t}_i) = \text{Cov}\{\bar{y}_{\cdot\cdot}\,(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\}$$

$$= \text{Cov}(\bar{y}_{\cdot\cdot}\ \bar{y}_{i\cdot}) - \text{var}\,(\bar{y}_{\cdot\cdot}) = \frac{r\sigma^2}{kr.r} - \frac{\sigma^2}{kr} = 0$$

Also $\text{cov}\,(\hat{t}_i, \hat{b}_j) = \text{cov}\{(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\,(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})\}$

$$= \text{Cov}\,(\bar{y}_{i\cdot}\ \bar{y}_{\cdot j}) - \text{Cov}\,(\bar{y}_{i\cdot}\ \bar{y}_{\cdot\cdot}) - \text{Cov}\,(\bar{y}_{\cdot\cdot}\ \bar{y}_{\cdot j}) + \text{var}\,(\bar{y}_{\cdot\cdot})$$

$$= \frac{\sigma^2}{kr} - \frac{r\sigma^2}{kr.r} - \frac{k\sigma^2}{kkr} + \frac{\sigma^2}{kr} = 0$$

Similarly the covariance between other combinations of the estimates can be shown to be zero showing that the estimates are mutually independent.

The total S.S. in this case, can be partitioned into three components as follows :

$$\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$$

$$= \sum_i \sum_j \{(\overline{y_i.} - \overline{y..}) + (\overline{y._j} - \overline{y..}) + (y_{ij} - \overline{y_i.} - \overline{y._j} + \overline{y..})\}^2$$

$$= r\sum_i (\overline{y_i.} - \overline{y..})^2 + k\sum_j (\overline{y._j} - \overline{y}..)^2 + \sum_i \sum_j (y_{ij} - \overline{y_i.} - \overline{y._j} + \overline{y}..)^2$$

all other product terms vanish,

Thus we have, Total S. S = Treatment S.S.+ Block S.S. + Error S.S.

Now, we have to show that different components of sum of squares follows χ^2-distribution with appropriate degrees of freedom.

We know, $y_{ij} = \mu + t_i + b_j + e_{ij}$

$$\overline{y_i.} = \mu + t_i + \overline{b} + \overline{e_i.}$$

$$\overline{y_j} = \mu + \overline{t} + b_j + \overline{e._j}$$

$$y.. = \mu + \overline{t} + \overline{b} + \overline{e}..$$

Now, $r \sum_i (\overline{y_i.} - \overline{y}..)^2 = r\sum_i (t_i - \overline{t} + \overline{e_i.} - \overline{e}..)^2$

$= r\sum_i (t'_i + \overline{e_i.} - \overline{e}..)^2$; considering $t_i - \overline{t} = t_i'$.

Expanding R.H.S, taking expectation on both the sides and assuming $t'_i = 0$ under $H_0 : t_1 = t_2 = ... = t_k$ we have, $E[r\sum(\overline{y_i.} - \overline{y}..)^2] = (k-1)\sigma^2$

or, $E\left[\dfrac{\sum(\overline{y_i.} - \overline{y}..)^2}{\sigma^2/r}\right] = (k-1)$, which implies that $\dfrac{\sum(\overline{y_i.} - \overline{y}..)^2}{\sigma^2/r}$

is distributed as χ^2 with (k - 1) d. f.

Similarly, it can be shown that

$E\left[\dfrac{\sum(\overline{y_j.} - \overline{y}..)^2}{\sigma^2/k}\right] = (r-1)$, indicating that $\left[\dfrac{\sum(\overline{y_j.} - \overline{y}..)^2}{\sigma^2/k}\right]$

is distributed as χ^2 with (r - 1) d. f.

Now the Error S.S. $= \sum_i \sum_j (y_{ij} - \overline{y_i.} - \overline{y._j} + \overline{y}..)^2$

$$= \sum_i \sum_j (e_{ij} - \overline{e_i.} - \overline{e._j} + \overline{e..})^2$$

$$= \sum_i\sum_j(e_{ij}^2 + \overline{e_i}.^2 + \overline{e_{.j}}^2 + \overline{e}..^2 + 2e_{ij}\ \overline{e}.. - 2\ \overline{e_i}.\ \overline{e}.. - 2\ \overline{e_{.j}}\ \overline{e}..)$$

$$-2e_{ij}\ \overline{e_{.j}} + 2\ \overline{e_i}.\ \overline{e_{.j}}. - 2e_{ij}\ \overline{e_i}.)$$

Taking expectation on both the sides we have the R. H. S. as follows :

$$Kr\sigma^2 + \frac{kr\sigma^2}{r} + \frac{kr\sigma^2}{k} + \frac{kr^2}{kr} + \frac{2kr\sigma^2}{kr} - \frac{2krr\,\sigma^2}{krr}$$

$$\frac{-\ 2krk\sigma^2}{krk} - \frac{2kr\sigma^2}{k} + \frac{2kr\sigma^2}{kr} - \frac{2kr\sigma^2}{r} = \sigma^2\ (kr - k - r + 1) = \sigma^2\ (k - 1)\ (r - 1)$$

Therefore, $E \sum_i\sum_j(y_{ij} - \overline{y_i}. - \overline{y_{.j}} + \overline{y..})^2/\sigma^2 = (k - 1)\ (r - 1)$

which implies that $\sum_i\sum_j(y_{ij} - \overline{y_i}. - \overline{y_{.j}}. + \overline{y}..)^2/\sigma^2$ is

distributed as χ^2 with $(k - 1)\ (r - 1)$ d. f.

From the additive property of χ^2 it can be said that $\sum_i\sum_j \frac{(y_{ij} - \overline{y}..)^2}{\sigma^2}$ is also

distributed as χ^2 with $(kr - 1)$ d. f. It can be shown independently also.

Thus it is seen that each of the components of sum of squares is independently distributed as χ^2 with appropriate d. f.

Now considering $H_0 : t_1 = t_2 = \ldots\ldots = t_k$, we have the test criterion,

$$F = \frac{r\sum(\overline{y_i}. - \overline{y}..)^2/(k - 1)}{\sum_i\sum_j(y_{ij} - \overline{y_i}. - \overline{y_{.j}} + \overline{y..})^2/(r - 1)\ (k - 1)}$$ which is distributed as F

with $(k - 1)$ and $(r - 1)\ (k - 1)$ d. f.

Again considering $H_0 : b_1 = b_2 = \ldots\ldots = b_r$, we have the test criterion,

$$F = \frac{r\sum(\overline{y_{.j}} - \overline{y})^2/(r - 1)}{\sum_i\sum_j(y_{ij} - \overline{y_i}. - \overline{y_{.j}} + \overline{y..})^2/(r - 1)\ (k - 1)}$$ which is distributed as F

with $(r - 1)$ and $(r - 1)\ (k - 1)$ d. f.

Method of calculations of different sum of squares :

Correction factor (C. F.) $= \frac{y_{..}^2}{rk}$.

Total sum of squares $= \Sigma\Sigma y_{ij}^2 - C.F. = T_o$, say.

Treatment sum of squares $= \frac{\Sigma y_{i.}^2}{r} - C.F. = T$, say.

Block sum of squares $= \frac{\Sigma y_{.j}^2}{k} - C.F. = B$ say.

Error sum of squares is obtained by usual subtraction i.e.

Error S.S. $= T_o - T - B = E$, say.

Now the analysis of variance table for testing null hypothesis,

H_0 : The effects of all the treatments are same, is as follows.

Table-11.3
ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F
Treatment	k - 1	T	$T' = T/_{(k-1)}$	$\frac{T'}{E'} = F_1$
Block	r - 1	B	$B' = B/_{(r-1)}$	$\frac{B'}{E'} = F_2$
Error	(k - 1) (r - 1)	E	$E' = \frac{E}{(r-1)(k-1)}$	
Total	rk - 1	T_o		

If the calculated value of F_1 with (k - 1) and (k - 1) (r - 1) d. f. is greater than the tabulated value of F with same d. f. and at 100 α % level of significance, then the hypothesis may be rejected otherwise the hypothesis may be accepted.

Similar hypothesis may be considered for block effects and a conclusion can be drawn with the help of F_2 also.

Example 11.2 Six different level of a certain fertiliser were tried in a randomised block design with 4 blocks at a certain agricultural farm to study the effects of the levels of fertiliser on cotton crop.

The yield per plot in kg for different levels of fertiliser and blocks are given systematically below for analysis.

Table-11.4
Cotton yield per plot in kg.

Levels of	Block			
Fertiliser	1	2	3	4
F_1	6.90	4.60	4.40	4.81
F_2	6.48	5.57	4.28	4.45
F_3	6.52	7.60	5.30	5.30
F_4	6.90	6.65	6.75	7.75
F_5	6.00	6.18	5.50	5.50
F_6	7.90	7.57	6.80	6.62

Solution : The block totals, treatment totals, and grand total are as follows :
Block totals : $y_{\cdot 1} = 40.70$, $y_{\cdot 2} = 38.17$, $y_{\cdot 3} = 33.03$, $y_{\cdot 4} = 34.43$.
Treatment totals :
$y_{1\cdot} = 20.71$, $y_{2\cdot} = 20.78$, $y_{3\cdot} = 24.72$, $y_{4\cdot} = 28.05$, $y_{5\cdot} = 23.18$, $y_{6\cdot} = 28.89$.

Grand total = $y_{\cdot\cdot} = 146.33$, Correction factor C. F. $= \dfrac{y_{\cdot\cdot}^2}{4 \times 6} = 892.19$

Now different sum of squares are as follows :
Total S.S. $= \sum\sum y_{ij}^2 -$ C. F. $= 920.78 - 892.19 = 28.59$.

Block S.S. $= \dfrac{\sum y_{\cdot j}^2}{k} -$ C. F. $= 898.31 - 892.19 = 6.12$.

Treatment S.S. $= \dfrac{\sum y_{i\cdot}^2}{r} -$ C. F. $= 907.68 - 892.19 = 15.44$ and

Error S.S. = 7.03.

H_0 : The effecs of all the treatments are same i.e. the effect of all levels of fertiliser are same.

Table-11.5
ANOVA TABLE

Source of variations	d.f.	S.S.	M.S.	F	1%F
Block	3	6.12	2.040	4.350	5.42
Treatment	5	15.44	3.088	6.584	4.56
Error	15	7.03			
Total	23	28.59			

The calculated value of F with (5,15) d.f. corresponding to treatment is greater than the tabulated value of F with same d. f. at 1% level of significance. Hence it is highly significant and the hypothesis may be rejected.

Missing Observations : For some uncontroled causes the observations in some of the plots in an experiment may be missing. In agricultural experiment crop may be damaged by animal or by misuse of pest etc. Again in animal experiment, some of the animals may die during the course of experiment. In these cases. the number of observations per treatment are not same and thus the orthogonality is destroyed.

The analysis of these data in this type of design may be carried out by estimating the missing observations in such a way that the error sum of squares is minimum or by the usual method of analysis of non-orthogonal data. But the latter case is combersome and hence we would proceed with analysis after the estimation of missing observations.

Estimation of Missing Observations and Analysis in R.B.D. :

(i) Single missing observation :

Let us suppose that in a R.B.D. with k treatments in r blocks, one observation is missing and that is say, x_1. Let T_i. B_j and G be the total of the ith treatment jth block and grand total respectively excluding the missing observation x_1 which occurs in the ith treatment and jth block.

The error sum of squares can be expressed in terms of x_1 considering terms independent of x_1 as C.

$$\text{Therefore, } S = C + x_1^2 - \frac{(T_i + x_1)^2}{r} - \frac{(B_j + x_1)^2}{k} + \frac{(G + x_1)^2}{kr}$$

$$\text{Now } \frac{dS}{dx_1} = 2x_1 - 2\frac{(T_i + x_1)}{r} - \frac{2(B_j + x_1)}{k} + \frac{2(G + x_1)}{kr} = 0$$

$$\text{Solving we get, } \hat{x}_1 = \frac{kT_i + rB_j - G}{(k-1)(r-1).}$$

Thus the single missing observation x_1 is estimated.

(ii) Two missing observations.

In the above R.B.D. if two observations x_1 and x_2 are missing, following are the possible cases to be considered.

(a) Two observations affecting different blocks and different treatments.

(b) Two observations affecting different blocks but same treatment.

(c) Two observations affecting same block but different treatments.

Case (a) : We assume that x_1 belongs to jth block and ith treatment and x_2 belongs to lth block and mth treatment. Let G be the grand total of the observations excluding x_1 and x_2. B_j and B_1 denote the total of the jth and lth blocks. T_i and T_m denote the total of the ith and mth treatments. The error sum of squares S can be expressed in terms of x_1 and x_2 and the remaining terms as C ;

$$S = C + x_1^2 + x_2^2 - \frac{(T_i + x_1)^2}{r} - \frac{(T_m + x_2)^2}{r} - \frac{(B_j + x_1)^2}{k} - \frac{(B_1 + x_2)^2}{k} + \frac{(G + x_1 + x_2)^2}{kr}$$

Now, $\frac{dS}{dx_1} = 0$ and $\frac{dS}{dx_2} = 0$ reduce to respectively.

$$x_1 - \frac{(T_1 + x_1)}{r} - \frac{(B_j + x_1)}{k} + \frac{(G + x_1 + x_2)}{kr} = 0$$

and $$x_2 - \frac{(T_m + x_2)}{r} - \frac{(B_1 + x_2)}{k} + \frac{(G + x_1 + x_2)}{kr} = 0.$$

Solving these equations we have,

$$\hat{x}_1 = \frac{(k-1)(r-1)(kT_i + rB_j - G) - (kT_m + rB_1 - G)}{(k-1)^2(r-1)^2 - 1}$$

and $$\hat{x}_2 = \frac{(k-1)(r-1)(kT_m + rB_1 - G) - (kT_i + rB_j - G)}{(k-1)^2(r-1)^2 - 1}.$$

Case (b) : In this case, the definition of block totals B_j and B_1 and grand total G remain same as in case (a). But here T_i is the total of the ith treatment in which two observations x_1 and x_2 are missing. In this case the error sum of squares can be written as,

$$S = C + x_1^2 + x_2^2 - \frac{(T_i + x_1 + x_2)2}{r} - \frac{(B_j + x_1)^2}{k} - \frac{(B_1 + x_2)^2}{k} + \frac{(G + x_1 + x_2)^2}{kr}$$

Now, $\frac{dS}{dx_1} = 0$ and $\frac{dS}{dx_2} = 0$ reduce $(k-1)(r-1)x_1 - (k-1)x_2 = kT_i + rBj - G$

and $-(k-1)x_1 + (k-1)(r-1)x_2 = kT_i + rB_1 - G.$ respectively.

Solving these two equations we have, $\hat{x}_1 = \frac{kT_i + (r-1)B_j + B_1 - G.}{(k-1)(r-2)}$

and $$\hat{x}_2 = \frac{kT_i + B_j + (r-1)B_1 - G}{(k-1)(r-2)}.$$

Case (c) : In this case the definition of treatment totals T_i and T_m and grand total G remain same as in case (a). But B_j denotes the total of the jth block in which both the missing observation x_1 and x_2 are lying. In this case the error sum of sqaures can be written as,

$$S = C + x_1^2 + x_2^2 - \frac{(T_i + x_1)^2}{r} - \frac{(T_m + x_2)^2}{r} - \frac{(B_j + x_1 + x_2)^2}{k} + \frac{(G + x_1 + x_2)^2}{kr}.$$

Now, $\frac{dS}{dx_1} = 0$ and $\frac{dS}{dx_2} = 0$ reduce to $(k-1)(r-1)x_1 - (r-1)x_2 = kT_i + rB_j - G$

and $-(r-1)x_1 + (k-1)(r-1)x_2 = kT_m + rB_j - G.$ respectively.

Solving we get, $\hat{x}_1 = \frac{(k-1)T_i + T_m + rB_j - G.}{(r-1)(k-2)}$ and

$$\hat{x}_2 = \frac{T_i + (k-1)T_m + rB_j - G}{(r-1)(k-2)}.$$

Thus the estimates of two missing observations in all possible cases are obtained.

There is another method of getting missing observations known as "iteration method" given by Yates (1938) which takes a lot of time and is subjected to contain larger bias. For more than two missing observations, reader is referred to Glenn and Kramer (1958).

The method of analysis in case of missing observations is as follows :

Table-11.6

ANOVA TABLE

Source of variation	Method of calculating sum of squares	d.f.
(i) Total	Original data	(kr - 1) - p*
(ii) Error	Completed data	(k - 1)(r - 1)-p*
(iii) Block + treatment	(i) - (ii)	k + r - 2
(iv) Block	Original data	(r - 1)
(v) Treatment	(iii) - (iv)	(k - 1)

p* in the components of d. f. indicates number of missing observations.

Example 11.3 : A Randomised block design with 4 varieties of paddy conducted in 5 blocks gave the following yield/acre in which two observations were missing. Estimate the missing observations and carry out the analysis of variance and draw conclusion over the effects of treatment i.e. paddy varieties.

Table-11.7

Block	Varieties			
	A	B	C	D
1	44.5	46.6	41.3	34.1
2	48.0	*	40.3	34.0
3	52.1	44.9	40.1	33.3
4	50.0	45.0	35.1	*
5	48.0	50.2	46.1	35.6

Solution : We know,

$$\hat{x}_1 = \frac{(r-1)(k-1)(kT_i + rB_j - G) - (kT_m + rB_1 - G)}{(k-1)^2 (r-1)^2 - 1}$$

$$\hat{x}_2 = \frac{(r-1)(k-1)(kT_m + rB_1 - G) - (kT_i + rB_j - G)}{(k-1)^2 (r-1)^2 - 1}$$

where the notations have their usual meaning.

Here $T_{i.} = 186.7$, $Bj = 122.3$

$T_m = 137.0$, $B_1 = 130.1$,

$G = 769.2$, $r = 5$, $k = 4$.

$$\therefore \hat{x}_1 = \frac{4 \times 3 \times 589.1 - 429.3}{4^2 \times 3^2 - 1} = \frac{6639.9}{143} = 46.43.$$

$$\text{and } \hat{x}_2 = \frac{4 \times 3 \times 429.3 - 589.1}{4^2 \times 3^2 - 1} = \frac{4562.5}{143} = 31.91.$$

Now different components of sum of squares :

$$\text{C.F. (original data)} = \frac{769.2^2}{18} = 32870.48.$$

$$\text{C. F. (completed data)} = \frac{847.54^2}{20} = 35916.2$$

Total S. S. (original data) = 33525.34 - 32870.48 = 654.86.

Total S. S. (completed data) = 36699.997-35916.2=783.173.

Block S.S. (completed data) = 35959.86 - 35916.2 = 43.66.

Treatment S.S. (completed data) = 36629.24 - 35916.20 = 713.04.

Error S.S. (completed data) = 783.17 - 43.66 - 713.04 = 26.47.

$$\text{Block S.S. (original data)} = \frac{166.5^2}{4} + \frac{122.8^2}{3} + \frac{170.4^2}{4} + \frac{130.1^2}{3}$$

$$+ \frac{179.9^2}{4} - 32870.48 = 37.89.$$

H_0 : The effects of all the treatments are equal.

Table-11.8
ANOVA TABLE

Source of variation	d.f.	Method of calculating	S.S.	M.S.	F	%F
(i) Total	17	Original data	654.86			
(ii) Error	10	Completed data	26.46	2.65		
(iii) Block + treatment	7	(i) - (ii)	628.39			
(iv) Block	4	Original data	37.89			
(v) Treatment	3	(iii) - (iv)	590.50	196.84	74.24	6.55

Here the calculated value of F with (3, 10) d. f. is greater than the theoretical value of F at 1% level of significance, therefore, the calculated value of F is highly significant and the hypothesis may be rejected.

R. B.D. with multiple observations made in each plot per block :

We may have to face some situations where single observation in each plot per block is not desirable where sampling is adopted to choose a sampling unit to obtain data that can provide necessary information.

For simplicity sake, we consider a constant number of observations, say, s observations made in each plot. There are k treatments each replicated in r blocks. The model can be written as,

$$y_{ijp} = \mu + t_i + b_j + (tb)_{ij} + e_{ijp}$$

where y_{ijp} is the observations on the pth sample for ith treatment in the jth block. ($i = 1, 2, \ldots\ldots k$; $j = 1, 2, \ldots\ldots r$ and $p = 1, 2, \ldots\ldots s$).

μ is the general mean,
t_i is the ith treatment effect.
b_j is the jth block effect.

$(tb)_{ij}$ is the interaction between treatment and block.

e_{ijp}, the sampling error which are normally and independently distributed with 0 mean and variance σ^2.

The estimation of different parameters and partitioning of the total sum of squares into different components can be performed as usual.

The calculation of different sum of squares due to different components of ANOVA TABLE can be obtained as follows :

Grand total = $y_{...} = \sum_i \sum_j \sum_p y_{ijp}$, Correction Factor (C. F.) = $\frac{y_{...}^2}{rks}$

Total S.S. $= \sum_i \sum_j \sum_p y^2_{ijp}$ - C. F. = T_o, say.

A two-way table like Treatment x Block is to be prepared for the calculation of the following components. The cell totals being y_{ij}.

Total S.S. (from Treatment x Block table) = $\sum_j \sum_j \frac{y_{ij}.^2}{s}$ - C. F. = T_t, say.

Treatment S.S. $= \frac{\sum y_i.^2.}{rs}$ - C. F. = T

Block S.S. $= \frac{\sum y_{.j}.^2}{ks}$ - C.F. = B.

Interaction between treatment and Block S.S. = T_t - T - B = I.

S.S. due to sampling error = T_o - T - B - I = E.

To test the null hypothesis H_0 : The treatment effects are equal, the ANOVA TABLE can be prepared as given in Table -11.9.

Table-11.9
ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F
Treatment	(k - 1)	T	$T' = \frac{T}{(k-1)}$	T'/E'
Block	(r - 1)	B		
Block x Treatment	(k - 1)(r - 1)	I		
Sampling error	rk(s - 1)	E	$E' = \frac{E}{rk(s-1)}$	
Total	krs - 1	T_o		

The conclusion can be drawn as usual.

Example 11.4 To study the effect of differences in the number of plants per hill on the growth of Maize crop, a randomised block design with 5 randomly selected cobs per plot was laid in 3 replications or blocks. The treatments are,

A-one plant/hill ; B - two plants / hill

C - there plants / hill ; D - four plants / hill.

The following table gave data on the length of cobs. Analyse the data and give your comment on the treatments.

	Treatment			
Replications	A	B	C	D
1	9.3	9.0	8.6	6.4
	8.8	9.0	7.0	7.2
	9.0	10.5	8.4	6.8
	8.8	8.9	9.1	7.7
	8.6	9.2	8.2	6.0
2	10.2	9.7	9.0	6.4
	9.0	10.0	8.0	7.4
	9.4	9.2	8.1	6.8
	9.6	10.5	8.2	6.8
	9.8	10.3	7.0	6.6
3	9.9	8.4	7.5	6.3
	10.4	9.4	7.5	6.7
	11.0	8.2	8.5	6.0
	10.8	9.1	8.0	7.0
	10.0	9.8	8.6	7.3.

Solution : At first we prepare a two-way table of replication x treatment.

Table-11.10

	Treatment				Total
Replications	A	B	C	D	
1	44.5	46.6	41.3	34.1	166.5
2	48.0	49.7	40.3	34.0	172.0
3	52.1	44.9	40.1	33.3	170.4
Total	144.6	141.2	121.7	101.4	508.9

C.F. $= \frac{508.9^2}{60} = 4316.32.$

Total S.S. (of the two-way table) $= \frac{22021.81}{5}$ - C. F. = 88.04

Replication S. S. $= \frac{86342.41}{20}$ - C. F. = 4317.12 - C. F. = 0.80.

Treatments S.S. $= \frac{65939.45}{15}$ - C.F. = 4395.96 - C.F. = 79.64.

Rep. x treat. S.S. (per pot error) = 88.04 - 0.80 - 79.64 = 7.6

Total S.S. (from entire data) = 4420.01 - C.F. = 103.69.

S.S. due to sampling error = 103.69 - 7.6 - 79.64 - 0.80 = 15.65.

H_0 : The effects of all the treatments are same.

Table-11.11

ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F	1%F
Replication	2	0.80	0.40		
Treatment	3	79.65	26.55	80.45	4.284
Rep X treat. (per plot error)	6	7.6	1.27		
Sampling error	48	15.65	0.33		
Total	59				

The calculated value of F is highly significant and the hypothesis may be rejected.

Latin Square Design (L.S.D.) :

In randomised block design, the experimental material is divided into groups of homogeneous units in one direction which increases the efficiency of the design rather than C.R.D. The latin square design is an improvement over R.B.D. obtained by classifying the experimental material in two directions rowwise and columnwise in such a way that the differences among rows and columns representing major sources of variation and they are orthogonal to each other. Though it is not necessary that the two factors should always be called row and column, it may be the levels of two factors also. Thus in a latin square of size v, the arrangement of v treatments in v^2 positions should be made in such a way that every row and every column contain every treatment precisely once.and make a perfect replication. Thus the error variance can be reduced considerably.

Latin square design is the most efficient design among the basic designs. The analysis is available for any member of missing observations.

The chief disadvantages that the number of rows, columns and treatments must be same i.e. the experimental unit must be a perfect squares which may not always be practical. The analysis depends on the assumption that the interaction between rows and columns is not present.

Layout : A standard square of required size is selected at random from the Tables for Statistician and Biometricians (Fishers and Yates 1948). All the columns are arranged after randomisation and similar randomisation is done for all rows except the first one to get the final lay-out of the latin square design.

Analysis : Let us consider a latin square of size v. For analysis of the data in this design, we consider the linear additive model,

$$y_{ijs} = \mu + r_i + c_j + t_s + e_{ijs}$$

where y_{ijs} is the observation of the sth treatment in the ith row and jth column ($i = 1, 2, \ldots\ldots v$, $j = 1, 2, \ldots\ldots v$: $s = 1, 2, \ldots\ldots v$).

μ is the general mean,

r_i is the effect due to ith row,

c_j is the effect due to jth column,

t_s is the effect due to sth treatment,

and e_{ijs}, the error components which are assumed to be independently and normally distributed with 0 mean and variance σ^2.

In the latin square v treatments are arranged in v rows and in v columns.

Let $y_{\ldots} = \sum_i \sum_j y_{ijs}$, grand total of observations.

$y_{i\cdot\cdot} = \sum_j y_{ijs}$, ith row total. $y_{\cdot j \cdot} = \sum_i y_{ijs}$, jth column total.

$y_{\cdot\cdot s} = \sum_s y_{ijs}$, sth treatment total

The least square estimate of μ, r_i. c_j and t_s can be obtained by minimising the error sum of squares denoted by $S = \sum_i \sum_j e_{ijs}^2 = \sum_i \sum_j [y_{ijs} - \mu - r_i - c_j - t_s]^2$.

In this case, we get four normal equations which can be solved after

imposing the restrictions $\sum_i r_i = \sum_j c_j = \sum_s t_s = 0$ and the estimates are,

$$\hat{\mu} = \overline{y}_{\ldots} \text{ where } \overline{y}_{\ldots} = \text{grand mean} = \frac{y_{\ldots}}{v^2}.$$

$$\hat{r}_i = \overline{y_{i}}_{\cdot\cdot} - \overline{y}_{\ldots} \text{ where } \overline{y}_{i\cdot\cdot} = \frac{y_{i\cdot\cdot}}{v}$$

$$\hat{c}_j = \overline{y_{\cdot j \cdot}} - \overline{y}_{\ldots} \text{ where } \overline{y}_{\cdot j \cdot} = \frac{y_{\cdot j \cdot}}{v}. \text{ and}$$

$\hat{t}_s = \overline{y}_{..s} - \overline{y}_{...}$ where $\overline{y}_{..s} = \frac{y_{..s}}{v}$

To show that the estimates are independent

We have, $Cov(\hat{\mu}, \hat{r}_i) = Cov\{\overline{y}_{...}(\overline{y}_{i..} - \overline{y}_{...})\}$

$= Cov(\overline{y}_{...}\ \overline{y}_{i..}) - Var(\overline{y}_{...})$

$= \frac{v\sigma^2}{v^2 v} - \frac{\sigma^2}{v^2} = 0$. Hence $\hat{\mu}$ and $\hat{r}_i$ are independent.

Now, $Cov(\hat{r}_i \hat{c}_j) = Cov\{(\overline{y}_{i..} - \overline{y}_{...})(\overline{y_{.j.}} - \overline{y}_{...})\}$

$= Cov(\overline{y}_{i..}\ \overline{y}_{.j.}) - Cov(\overline{y}_{i..}\ \overline{y}_{...}) - Cov(\overline{y}_{.j.}\ \overline{y}_{...}) + Var(\overline{y}_{...})$

$= \frac{\sigma^2}{vv} - \frac{v\sigma^2}{vv^2} - \frac{v\sigma^2}{vv^2} + \frac{\sigma^2}{v^2} = 0$. Hence $\hat{r}_i$ and $\hat{c}_j$ are independent.

Similarly it can be shown that the covariances between all possible pairs of estimates are zero, indicating that the estimates are mutually independent.

The total S.S, in this case, can be partitioned into four components as follows:

$$\Sigma\Sigma(y_{ijs} - \overline{y}_{...})^2 = \sum_i\sum_j (y_{ijs} - \overline{y}_{i..} + \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{.j.} - \overline{y}_{..s} + \overline{y}_{..s} - \overline{y}_{...})^2$$

$$= \sum_i\sum_j \{(\overline{y}_{i..} - \overline{y}_{...}) + (\overline{y}_{.j.} - \overline{y}_{...}) + (\overline{y}_{..s} - \overline{y}_{...})$$

$$+ (y_{ijs} - \overline{y}_{i..} - \overline{y}_{.j.} - \overline{y}_{..s} + 2\overline{y}_{...})\}^2$$

$$= v\sum_i(\overline{y}_{i..} - \overline{y}_{...})^2 + v\sum_j(\overline{y}_{.j.} - \overline{y}_{...})^2 + v\sum_s(\overline{y}_{..s} - \overline{y}_{...})^2$$

$$+ \sum_i\sum_j(y_{ijs} - \overline{y}_{i..} - \overline{y}_{.j.} - \overline{y}_{..s} + 2\overline{y}_{...})^2$$ all other product terms vanish.

Thus we have, Total S.S. = Row S.S. + Col. S.S. + Treat. S.S. + Error S.S.

Now, we are to show that different components of sum of squares follow χ^2 - distribution with appropriate degrees of freedom.

We know,

$y_{ijs} = \mu + r_i + c_j + t_s + e_{ijs}$

$\overline{y}_{i..} = \mu + r_i + \overline{c} + \overline{t} + \overline{e}_{i..}$

$\overline{y}_{.j.} = \mu + \overline{r} + c_j + \overline{t} + \overline{e}_{.j.}$

$\overline{y}_{..s} = \mu + \overline{r} + \overline{c} + t_s + \overline{e}_{..s}$

$\overline{y}_{...} = \mu + \overline{r} + \overline{c} + \overline{t} + \overline{e}_{...}$

Also we know $(\overline{y}_{i..} - \overline{y}_{...}) = (r_i - \overline{r} + \overline{e}_{i..} - \overline{e}_{...})$

$= (r'_i + \overline{e}_{i..} - \overline{e}_{...})$, Putting $r_i - \overline{r} = r'_i$

$$\therefore\ v\sum_i(\overline{y}_{i..} - \overline{y}_{...})^2 = v\sum_i(r'_i + \overline{e}_{i..} - \overline{e}_{...})^2$$

Expanding R.H.S. taking expectation on both the sides and assuming $r'_i = 0$ under $H_0 : r_1 = r_2 = ... = r_v$, we have

$$E\left[v\sum_i(\overline{y}_{i..} - \overline{y}_{...})^2\right] = vE\sum_i \overline{e}_{i..}^2 + vE\sum_i \overline{e}_{...}^2 - 2vE\sum_i \overline{e}_{i..}\,\overline{e}_{...}$$

$$= \frac{v.v\,\sigma^2}{v} + \frac{v.v\sigma^2}{v^2} - \frac{2v^2.v\,\sigma^2}{v^2v} = v\sigma^2 - \sigma^2 = \sigma^2(v - 1).$$

$$\therefore\ E\left[\frac{\sum_i(\overline{y}_{i..} - \overline{y}_{...})^2}{\sigma^2/v}\right] = (v - 1)\text{ ; indicating that}$$

$$\frac{\sum_i(\overline{y}_{i..} - \overline{y}_{...})^2}{\sigma^2/v}$$ is distributed as χ^2 with $(v - 1)$ d.f.

Similarly Column S.S. and Treatment S.S. can be shown to be distributed as χ^2 with $(v - 1)$ independently.

Now, the error S.S. = $\sum_i\sum_j(e_{ijs} - \overline{e}_{i..} - \overline{e}_{.j.} - \overline{e}_{..s} + 2\,\overline{e}_{...})^2$

Expanding the R.H.S. and taking expectation on both the sides we have,

$$E\left[\sum_i\sum_i(y_{ijs} - \overline{y}_{i..} - \overline{y}_{..s} + 2\,\overline{y}_{...})^2\right] = \sigma^2 (v-1)(v-2)$$

$$\therefore E\left[\frac{\sum_i\sum_j(y_{ijs} - \overline{y}_{i..} - \overline{y}_{.j.} - \overline{y}_{..s} + 2\,\overline{y}_{...})^2}{\sigma^2}\right] = (v-1)(v-2).$$

Which indicates that $\dfrac{\sum_i\sum_j(y_{ijs} - \overline{y}_{i..} - \overline{y}_{.j.} - \overline{y}_{..s} + 2\,\overline{y}_{...})^2}{\sigma^2}$

is distributed as χ^2 with (v -1) (v -2) d.f.

From the additive property of χ^2 it can be said that the

Total S.S. = $\sum_i\sum_j \dfrac{(y_{ijs} - \overline{y}_{...})^2}{\sigma^2}$ is also distributed as χ^2 with $(v^2 - 1)$ d.f.

It can be shown independently also.

Thus it is seen that each of the components of sum of squares is independently distributed as χ^2 with appropriate d.f.

Now considering $H_0 : r_1 = r_2 = \ldots\ldots = r_v$, we have the test criterion,

$$F = \frac{v\sum(\overline{y}_{i..} - \overline{y}_{...})^2/(v-1)}{\sum_i\sum_j(y_{ijs} - \overline{y}_{i..} - \overline{y}_{.j.} - \overline{y}_{..s} + 2\,\overline{y}_{...})^2/(v-1)(v-2)}$$

which is distributed as F with (v -1) and (v -1) (v -2) d. f.

Again, considering $H_0 : c_1 = c_2 = \ldots\ldots = c_v$, we have the test criterion,

$$F = \frac{v\sum_i(\overline{y}_{.j.} - \overline{y}_{...})^2/(v-1)}{\sum_i\sum_j(y_{ijs} - \overline{y}_{i..} - \overline{y}_{.j.} - \overline{y}_{..s} + 2\,\overline{y}_{...})^2/(v-1)(v-2)}$$

which is distributed as F with (v - 1) and (v - 1) (v - 2) d. f.

Once again, considering $H_0 : t_1 = t_2 = \ldots\ldots = t_v$, we have the test criterion,

$$F = \frac{v\sum_{i}(\bar{y}_{..s} - \bar{y}_{...})^2/(v-1)}{\sum_{i}\sum_{j}(y_{ijs} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..s} + 2\bar{y}_{...})^2/(v-1)(v-2)}$$

which is distributed as F with (v - 1) and (v - 1) (v - 2) d. f.

Method of calculations of different components of sum of squares are as follows :

Total S.S. $= \sum_{i}\sum_{j} y^2_{ijs} - \text{C.F.} = T_o$, say. where C. F. $= \frac{y_{...}^2}{v^2}$

Row. S.S. $= \frac{1}{v}\sum y_{i..}^2 - \text{C.F.} = R$, say. Column S.S. $= \frac{1}{v}\sum y^2_{.j.} - \text{C.F.} = C$, say.

Treatment S.S. $= \frac{1}{v}\sum y_{..s}^2 - \text{C.F.} = T$, say. Error S.S. $= T_o - R - C - T = E$, say.

Now the analysis of variance table for testing the null hypothesis

H_0 : The effects of all the treatments are same, can be furnished as given in Table 11.12.

TABLE-11.12
ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F
Treatment	(v -1)	T	$T' = \frac{T}{v-1}$	$\frac{T'}{E'}$
Row	(v -1)	R		
Column	(v - 1)	C		
Error	(v -1) (v -2)	E	$E' = \frac{E}{(v-1)(v-2)}$	
Total	v^2 -1	T_o		

For significant value of F, the hypothesis may be rejected otherwise the H_0 may be accepted.

Similar tests can be performed for testing hypothesis regarding column and row effects also.

Example 11.5 An experiment on cotton was conducted to study the effect of application of urea in combination with insecticidal sprays on the cotton yields. The lay-out of the latin square plan and yields of cotton per plot are given in Table 11.13. The rows of the table indicates the six different levels of moisture contents of soil and columns indicate the six different levels of spacing and $T_1, T_2, \ldots\ldots T_6$ indicate 6 different treatments obtained by taking some of the levels of urea and some levels of insecticides.

Table-11.13
Yields of Cotton/Plot

T_3 - 3.10	T_6 - 5.95	T_1 - 1.75	T_5 - 6.40	T_2 - 3.85	T_4 - 5.30
T_2 - 4.80	T_1 - 2.70	T_3 - 3.30	T_6 - 5.95	T_4 - 3.70	T_5 - 5.40
T_1 - 3.00	T_2 - 2.95	T_5 - 6.70	T_4 - 5.95	T_6 - 7.75	T_3 - 7.10
T_5 - 6.40	T_4 - 5.80	T_2 - 3.80	T_3 - 6.55	T_1 - 4.80	T_6 - 9.40
T_6 - 5.20	T_3 - 4.85	T_4 - 6.60	T_2 - 4.60	T_5 - 7.00	T_1 - 5.00
T_4 - 4.25	T_5 - 6.65	T_6 - 9.30	T_1 - 4.95	T_3 - 9.30	T_2 - 8.40

Analyse the data and give your comments.

Solution :

Row Totals :	$y_{1\cdot\cdot}$	$y_{2\cdot\cdot}$	$y_{3\cdot\cdot}$	$y_{4\cdot\cdot}$	$y_{5\cdot\cdot}$	$y_{6\cdot\cdot}$
	26.35	25.85	33.45	36.75	33.25	42.85
Column Totals :	$y_{\cdot 1\cdot}$	$y_{\cdot 2\cdot}$	$y_{\cdot 3\cdot}$	$y_{\cdot 4\cdot}$	$y_{\cdot 5\cdot}$	$y_{\cdot 6\cdot}$
	26.75	28.90	31.45	34.40	36.40	40.60
Treatment Totals :	$y_{\cdot\cdot 1}$	$y_{\cdot 2\cdot}$	$y_{\cdot\cdot 3}$	$y_{\cdot\cdot 4}$	$y_{\cdot\cdot 5}$	$y_{\cdot 6\cdot}$
	22.20	28.40	34.20	31.60	38.55	43.55

$$\text{Correction factor (C. F.)} = \frac{y_{\cdots}^2}{v^2} = \frac{198.5^2}{36} = 1094.51$$

$$\text{Total S.S.} = \Sigma\Sigma y^2_{ijs} - \text{C.F.} = 1222.84 - 1094.51 = 128.33$$

$$\text{Row S.S.} = \frac{1}{v}\Sigma y_{i\cdot\cdot}^2 - \text{C.F.} = \frac{6773.695}{6} - 1094.51 = 34.44.$$

$$\text{Column S.S.} = \frac{1}{v}\Sigma y^2{\cdot j\cdot} - \text{C.F.} = \frac{6696.555}{6} - 1094.51 = 21.58.$$

Treatment S.S. $= \frac{1}{v}\Sigma y^2..s - \text{C.F.} = \frac{6850.305}{6} - 1094.51 = 47.21$

Error S.S. = 128.33 - 34.44 - 21.58 - 47.21 = 25.10.

H_0 : The effects of all the treatments are equal.

Table-11.14

ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F	1%F
Row	5	34.44	6.888		
Column	5	21.58	4.316		
Treatment	5	47.21	9.442	7.523	4.10
Error	20	25.10	1.255		
Total	35	128.33			

The calculated value of F is highly significant and therefore, the hypothesis may be rejected.

Estimation of Missing Observations and Analysis in Latin Square Design :

(i) **Single missing observation :** Let there be one missing observation, denoted by x. Let R_i, C_j, T_s and G be the total of the ith row, jth column, sth treatment and grand total respectively obtained from the original data where one observation is missing. The error S.S. can be expressed in terms of x and taking other quantities as C,

$$\text{Error S.S.} = S = C + x^2 - \frac{(R_i + x)^2}{v} - \frac{(C_j + x)^2}{v} - \frac{(T_s + x)^2}{v} + \frac{2(G + x)^2}{v^2}$$

Differentiating S with respect to x and equating to zero we have after simplification ; $\hat{x} = \frac{v(R_i + C_j + t_s) - 2G}{(v-1)(v-2)}$

Thus the single missing observation x is estimated.

(ii) **Two missing observations :** In a latin square design of order v X v if two observations x_1 and x_2 are missing, following are the possible cases to be cosidered according to Shil and Debnath (1986).

(a) Missing observations are in different rows and columns affecting different treatments.

(b) Missing observations are in different rows and columns affecting same treatment.

(c) Missing observations are in same row but in different columns affecting necessarily different treatments.

(d) Missing observations are in different rows but in same column affecting necessarily different treatments.

Case (a) : Let us consider two missing observations x_1 and x_2 in a latin square design with v rows, v columns and v treatments. Let Ts and T's be the total of the sth and s'th treatments without considering the missing observations x_1 and x_2 respectively. Similarly R_i and R'_i the row totals and C_j and C'_j, the column totals can be defined. Let G be grand total of all the observations without considering x_1 and x_2. The error sum of squares (S) can be written below in terms of x_1 and x_2 and all other terms as C,

$$S = C + x_1^2 + x_2^2 - \frac{(T_s + x_1)^2}{v} - \frac{(T'_s + x_2)^2}{v} - \frac{(R_i + x_1)^2}{v} - \frac{(R'_i + x_2)^2}{v}$$

$$- \frac{(C_j + x_1)^2}{v} - \frac{(C'_j + x_2)^2}{v} + \frac{2(G + x_1 + x_2)^2}{v^2}$$

Now, $\frac{dS}{dx_1} = 0$ and $\frac{dS}{dx_2} = 0$ reduce to

$$(v - 1)(v - 2) x_1 + 2x_2 = v(T_s + R_i + C_j) - 2G$$

$$2x_1 + (v - 1)(v - 2)x_2 = v(T'_s + R'_i + C'_j) - 2G.$$

Solving these two equations, the estimates of x_1 and x_2 can be obtained as follows :

$$\hat{x}_1 = \frac{1}{(k - 3)(k^2 - 3k + 4)} [(k - 1)(k - 2)(T_s + R_i + C_j) - 2(T'_s + R'_i + C'_j) - 2(k - 3)G]$$

$$\hat{x}_2 = \frac{1}{(k - 3)(k^2 - 3k + 4)} [(k - 1)(k - 2)(T'_s + R'_i + C'_j) - 2(T_s + R_i + C_j) - 2(k - 3)G]$$

Case (b) : In this case, the error sum of squares, can be written as,

$$S = C + x_1^2 + x_2^2 - \frac{(T_s + x_1 + x_2)^2}{v} - \frac{(R_i + x_1)^2}{v} - \frac{(R_i' + x_2)^2}{v}$$

$$- \frac{(C_j + x_1)^2}{v} - \frac{(C'j + x_2)^2}{v} + \frac{2(G + x_1 + x_2)^2}{v^2}$$

Explanations of all the terms here are same as in case (a) except that of Ts, which indicates the total of sth treatment in which the observations x_1 and x_2 are missing. Proceeding as in case (a) we have the estimates of x_1 and x_2 as follows :

$$\hat{x}_1 = \frac{1}{(v - 2)^2} [vT_s + (v - 1)(R_i + C_j) + R'_i + C'_j - 2G]$$

and $$\hat{x}_2 = \frac{1}{(v - 2)^2} [vT_s + (v - 1)(R_i' + C'_j) + R_i + C_j - 2G]$$

Case (c) : In this case, the error sum of squares can be written as,

$$S = C + x_1^2 + x_2^2 - \frac{(Ts + x_1)^2}{v} - \frac{(T's + x_2)^2}{v} - \frac{(R_i + x_1 + x_2)^2}{v}$$

$$- \frac{(C_j + x_1)^2}{v} - \frac{(C'_j + x_2)^2}{v} + \frac{2(G + x_1 + x_2)^2}{v^2}.$$

Explanation of all the terms in S are same as in case (a) except that of R_i which indicates the ith row total in which two observations x_1 and x_2 are missing.

Proceeding as in case (a) we have the estimates of x_1 and x_2 as follows.

$$\hat{x}_1 = \frac{1}{(v-2)^2}[(v-1)(Ts + C_j) + v R_i + Ts' + C_j' - 2G] \text{ and}$$

$$\hat{x}_2 = \frac{1}{(v-2)^2}[(v-1)(T's + C_j') + vR_i + Ts + C_j - 2G]$$

Case (d). The error sum of squares can be written as,

$$S = C + x_1^2 + x_2^2 - \frac{(Ts + x_1)^2}{v} - \frac{(Ts' + x_2)^2}{v} - \frac{(R_i + x_1)^2}{v} - \frac{(R_i' + x_2)^2}{v}$$

$$- \frac{(C_j + x_1 + x_2)^2}{v} + \frac{2(G + x_1 + x_2)^2}{v^2}.$$

The explanation of all the terms here are same as given in case (a) except that of C_j which indicates the total of the jth column in which both the observations x_1 and x_2 are missing. Proceeding as in case (a) we have the estimates of x_1 and x_2 as follows :

$$\hat{x}_1 = \frac{1}{(v-2)^2}[(v-1)(T_s + R_i) + vC_j + T_s' + R_i' - 2G] \text{ and}$$

$$\hat{x}_2 = \frac{1}{(v-2)^2}[(v-1)(T_s' + R_i') + vC_j + T_s + R_i - 2G]$$

Thus the estimates of two missing observations for all possible cases are obtained.

When more than two observations are missing the number of possbile cases increases rapidly and the estimation procedure becomes combersome. In that case we suggest Yates (1933) method of iteration.

The method of analysis, in case of missing observations : Corrected error sum of squares E can be obtained by the usual method after substituting the estimated missing observations but in this way the corrected treatment sum of squares cannot be obtained. To get the corrected treatment sum of squares we adopt the method of estimating the missing observations as given in randomised block design by considering the rows and columns of latin design ignoring the treatment classification. The error sum of squares E_1 is calculated from the completed data thus obtained. Then $E - E_1$ gives the corrected treatment sum of squares. The degrees of freedom (d.f.) of the error sum of squares is reduced by the number of missing observations.

A clear-cut method of analysis of variance of the above type of data can be pointed out as given in Table-11.15.

Table-11.15

Source of variation	Method of calculating sum of squaress	d.f
(i) Total	Original data	$v^2 - 1 - P^*$
(ii) Error	Completed data	$(v - 1)(v - 2) - P^*$
(iii) Treatment + Row + Column	(i) - (ii)	$3v - 3$
(iv) Row + Column	Original data	$\left.\begin{matrix} v - 1 \\ v - 1 \end{matrix}\right\}$
(v) Treatment	(iii) - (vi)	$v - 1$

P^* indicates the number of missing observations.

Example 11.6 Six different insecticidal sprays ($T_1, T_2 \ldots\ldots T_6$) on the cotton yields were applied in a latin square experiment in the following type of lay-out. Two observations were missing in the plan, the data were collected as follows :

Table-11.16

T_3 3.10	T_6 5.95	T_1 1.75	T_5 6.49	T_2 3.85	T_4 5.30
T_2 4.80	T_1 2.70	T_3 3.30	T_6 5.95	T_4 3.70	T_5 5.40
T_1 3.00	T_2 2.95	T_5 *	T_4 5.95	T_6 7.75	T_3 7.10
T_5 6.40	T_4 5.80	T_2 3.80	T_3 6.55	T_1 4.80	T_6 9.40
T_6 5.20	T_3 4.85	T_4 6.60	T_2 4.60	T_5 *	T_1 5.00
T_4 4.25	T_5 6.65	T_6 9.30	T_1 4.95	T_3 9.30	T_2 8.50

Estimate the missing values and analyse the data.

Solution : Since the missing values are in different rows, different columns but affecting the same treatment T_5, we have,

$$\hat{x}_1 = \frac{1}{(v-2)^2}\,[vTs + (v-1)(R_i + C_j) + R_i' + C_j' - 2G] \text{ and}$$

$$\hat{x}_2 = \frac{1}{(v-2)^2}\,[vTs + (v-1)(R_i' + C_j') + R_i + C_j - 2G]$$

where $\hat{x}_1$ and $\hat{x}_2$ are the estimated missing values in third row and third column and in the fifth row and fifth column respectively. All other symbols have the usual meaning expressed earlier.

Here, $G = 184.90$ $\quad Ri = 26.75$ $\quad C_j = 24.75$

$Ts = 24.95$ $\quad R'_i = 26.25$ $\quad C'_j = 29.40.$

Therefore we have, $\hat{x}_1 = 5.82$. and $\hat{x}_2 = 6.85$.

Correction factor (C. F.) (Original data) $= \frac{184.9^2}{34} = 1005.53,$

Correction factor (C.F.) (Completed data) $= \frac{197.57^2}{36} = 1084.28.$

Total S.S. (Original data) = 1130,64 - 1005.53 = 125.11.

Row S.S. (Original data) = 1040.57 - 1005.53 = 35.04

Col. S.S. (Original data) = 1027.15 - 1005.53 = 21.62.

Total S.S. (Completed data) = 1211.43 - 1084.28 = 127.15

Row S.S. (Completed data) = 1119.04 - 1084.28 = 34.76.

Col. S.S. (Completed data) = 1106.54 - 1084.28 = 22.26.

Treat.S.S (Completed data) = 1129.61 - 1084.28 = 45.33

Error S.S. (Completed data) = 24.80.

H_0 : Effects of all types of insicticidal sprays are equal.

Table-11.17

ANOVA TABLE.

Source of variation	Method of calculating sum of squares	d.f	S.S.	M.S.	F
(i) Total	Original data	33	125.11	1.38	
(ii) Error	Completed data	18	24.8		
(iii) Row + Col. + Treat.	(i) - (ii)	15	100.31		
(iv) Column	Original data	5	21.62		
Row	Original data	5	35.04		
(v) Treatment	(iii) - (iv).	5	43.65	8.73	6.33

The tabulated value of F with (5.18) d.f. at 1% level of significance is 4.25 which is smaller than the calculated value of F with same d.f. Therefore, the calculated value of F is highly significant and the hypothesis may be rejected.

Replicated Latin Square Design : When the number of treatments are 8 or more, latin square design should not be used because the number of replication are large and may not be available. On the other hand, a latin square design of order 2 x 2 cannot be adopted because in this case error d.f. cannot be obtained. For latin square of order 3 x 3, the error d.f. is 2 and for latin square of order 4 x 4, the error d.f. is 6. The error d. f. in both the above cases are not enough to give an effective analysis of variance. To increase the d.f. due to error in the above cases we repeat the experiment i.e. instead of taking one latin square, a number of say, p latin squares may be considered. The number of treatment in each of the p squares should be same and seperate randomisation is to be carried out in each case. The row and column classification should be maintained equal for all the squares. The design thus obtained is called replicated latin square design.

The analysis of data in this type of experiment is described as below :

Firstly, each of the p latin squares is analysed seperately following the method given earlier. The corresponding sum of squares are then added. This gives pooled row, column, treatment and error sum of squares. The pooled row sum of squares is called between row within squares sum of squares and similarly for the other pooled sum of squares.

From each of p squares, the v treatment totals are obtained and arranged ir a square x treatment table of order p X v.

Let T_{ts} donote the totals of all observations of the sth treatment in the tth square. the square x treatment table is obtained with these T_{ts} totals. Let p_t denote the total of all the observations in the t th square (t = 1. 2.........p) and Ts denoted the total of observations of the sth treatment from all latin squares (s = 1, 2.........v). p_t and Ts are the marginal totals of the square x treatment table.

Next, the following sum of squares are obtained.

$$\text{Correction factor (C. F.)} = \frac{\left(\sum_t p_t\right)^2}{pv^2}$$

$$\text{Sum of squares due to squares} = \frac{\Sigma pt^2}{v^2} - \text{C. F.}$$

$$\text{Sum of squares due to treatment} = \frac{\Sigma Ts^2}{vp} - \text{C. F.}$$

Sum of squares due to interaction treatment x square = Pooled treatment sum of squares - $\left(\frac{\Sigma Ts^2}{vp} - \text{C. F.}\right)$.

$$\text{Total sum of squares} = \sum_i \sum_j y_{ijst}^2 - \text{C. F.}$$

where y_{ijst} denotes the observation from the t th squares in its ith row, jth column and under sth treatment.

The partitioning of d.f. in the analysis of variance of data in replicated latin squares is shown in Table-11.18. The null hypothesis considered usually is ,

H_0: The treatment effects are same.

Table-11.18

Source of variation	degrees of freedom (d.f)
Squares	p - 1
Row (Pooled)	p(v -1)
Column (Pooled)	p(v -1)
Treatment	(v -1)
Treat. x Sq. Interaction	(p 1) (v -1)
Error (Pooled)	p(v -1) (v -2)
Total	pv^2 - 1

The test of significance regarding the null hypothesis stating the equality of all row and column effects are to be carried out as usual.

Example 11.7 The following 4 x 4 latin square experiment was conducted to compare the effect of 4 spacing A, B, C, and D on the yield lb/acre of certain variety of paddy. The whole experiment was repeated 3 times. The lay-out were as follows :

4 rows indicates = 4 different doses of fertilisers

4 cols indicates = 4 different levels of irrigations.

A 231	B 280	C 285	D 289
B 284	A 246	D 283	C 271
C 275	D 282	A 258	B 258
D 259	C 271	B 289	A 275

L. Square—1

B 215	C 310	D 280	A 280
C 219	B 241	A 249	D 265
D 180	A 239	B 290	C 260
A 210	D 245	C 275	B 271

L. Square—2

C 225	D 254	A 251	B 271
D 218	C 231	B 231	A 275
A 231	B 249	C 263	D 295
B 241	A 231	D 273	C 266

L. Square—3

Analyse data and give your comment on spacing.

Solution :

Latin Square—1.

Row Totals :	R_1	R_2	R_3	R_4
	1085	1084	1073	1094
Column Totals :	C_1	C_2	C_3	C_4
	1049	1079	1115	1093
Treat. (Spacing) Totals:	T_a	T_b	T_c	T_d
	1010	1111	1102	1113

Correction factor (C.F.) $= \dfrac{(4336)^2}{16} = 1175056.$

Row S.S. $= \dfrac{1085^2 + ... + 1094^2}{4}$ - C.F. = 1175111.5 - C.F. = 55.5

Col. S.S. $= \dfrac{1049^2 + ... + 1093^2}{4}$ - C.F. = 1175629 - C.F. = 573.

Treat. S.S. $= \dfrac{1010^2 + -...... + 1113^2}{4}$ - C.F. =1176898.5 - C.F. = 1842.5.

Total S.S. = 231^2 + ...+ 275^2 - C.F. = 1179154 - C.F. = 4098.

Error S.S. = 1627.

Latin Square-2,

Row Totals	R_1	R_2	R_3	R_4
	1085	974.	969	1001.
Column Totals	C_1	C_2	C_3	C_4
	824	1035	1094	1076
Treat. (Spacing) Totals	T_a	T_b	T_c	T_d
	978	1017	1064.	970.

Correction factor (C.F.) $= \dfrac{(4029)^2}{16} = 1014552.6$

Row S.S. $= \dfrac{1085^2 + ... + 1001^2}{4}$ - C. F. = 1016-15.8 - C. F. = 2163.2

Column S. S. $= \dfrac{824^2 + ... + 1076^2}{4}$ - C. F. =1026203.3 - C. F. = 11650.7.

Treat. S.S. $= \dfrac{978^2 + ... + 970^2}{4}$ - C. F. =1015942.8 - C. F. = 1389.7.

Total S.S. = 1031805 - C. F. = 17252.4. Error S.S. = 2048.8.

Latin Square-3.

	R_1	R_2	R_3	R_4
Row Totals:	1001	955	1038.	1011.
	C_1	C_2	C_3	C_4
Column Totals:	915.	965	1018	1107.
	T_a	T_b	T_c	T_d.
Treat. (Spacing) Totals:	988	992	985	1040.

Correction factor (C.F.) = $\frac{(4005)^2}{16}$ = 1002501.6.

Row S. S. = $\frac{1001^2 + \ldots + 1011^2}{2}$ - C. F. = 1003397.8 - C. F. = 896.2

Column S. S. = $\frac{915^2 + \ldots + 1107^2}{4}$ - C. F. = 1007555.8 - C. F. = 5054.2.

Treat. S. S. = $\frac{988^2 + \ldots + 1040^2}{4}$ - C. F. = 1003008.3 - C. F. = 506.7.

Total S. S. = 1009737 - C. F. = 7235.4. Error S. S. = 778.3.

Row S. S. (Pooled) = 3114.9. Column S. S. (Pooled) = 17277.9.

Treatment S. S. (Pooled) = 3738.9. Error S. S. (Pooled) = 4454.1

Table-11.19

Square x Treatment Table

Traet→ Square	A	B .	C	D	Total
1	1010	1111	1102	1113	4336
2	978	1017	1064	970	4029
3	988	992	985	1040	4035
Total	2976	3120	3151	3123	12370

Correction factor (C. F.) = $\frac{12365^2}{48}$ = 3187852.1

S. S. due to square = $\frac{4336^2 + \ldots + 4005^2}{16}$ - C. F. = 4258

S. S. due to Treatment = $\frac{2976^2 + \ldots + 3123^2}{12}$ - C. F. = 1556.7.

Int. S. S. due to (Treat. x Square) = Treat. S. S. (Pooled) - S.S. due to Treatment = 3738.9 - 1556.7 = 2182.2.

H_0 : Effects of all the treatments are same.

Table-11.20
ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F
Square	2	4258		
Row (Pooled)	9	3114.9		
Column (Pooled)	9	17277.9		
Treatment	3	1556.7	518.9	2.097
Int. Treatment x Square	6	2182.2		
Error (Pooled)	18	4454.1	247.45	
Total	47			

The tabulated value of F with (3, 18) d.f. at 5% level of significance is 3.16 which is greater than the calculated value of F with same d.f. Hence the calculated value is insignificant and the hypothesis may be accepted.

11.3 Cross-over Design

In an agricultural experiment if an experimental unit is used for several treatments in a squence i.e. if different fertilisers are used on the same experimental unit or in an animal husbandary experiment if a cow is given several feeds in a sequence at different periods, say, in different lactation stages, then in all the cases, the effects of the treatments applied in one period may carry over to the next period. Therefore, the design in which different treatments are applied to the same experimental unit in different periods is called cross-over design. It looks like a replicated latin square and is particularly appropriate when the difference between the rows is almost same in all replicates. Even if the difference between the rows is assumed to be large, the cross-over design may be used for small experiment where few degrees of freedom are available for error.

In this type of design we have to consider two cases, namely,

i) when it is assumed that the residual effect is nil.

ii) when the residual effect exists.

Case (i) When the residual effect is nil : Let us consider that the number of treatment be t, each replicated r times and to satisfy the condition of the experiment each treatment occurs equally often in each period and on each unit, then the cross over design will have t x r columns. Each column represents a replicate or block in a randomised block design. The treatments are randomised within the replicate in such a way that each treatment occurs once in the replicate and r times in each row.

The design can be used with any number of treatments subject to the restriction that the number replicates must be a multiple of the number of treatments.

The spliting of degrees of freedom in ANOVA is as follows :

Table-11.21

Source of variation	d.f
Replicate (Column)	tr-1
Row	t-1
Treatment	t-1
Error	(t-1) (tr-2)

Let us consider an animal husbandary experiment to observe the effect of three feeds A, B and C on milk production applied to 6 cows in 3 different lactation stages. It is welknown that the first lactation stage is the best, second lactation stage is medium and the third lactation stages is the wrost in connection with the milk production. To satisfy the condition of the constructions, we consider a cow to represent a replication and 3 rows represent 3 lactation stages. The lay-out can be shown as follows :

Replications

	R_1	R_2	R_3	R_4	R_5	R_6
Row 1	A	C	A	B	B	C
Row 2	C	B	B	C	A	A
Row 3	B	A	C	A	C	B

Case (ii) When there exists residual effect : We have seen in the earlier lay-out that to a same cow, say cow 1 indicated by R_1 is given feed A in the first stage i. e. stage 1, feed C in the second stage i.e. stage 2 and B in stage 3 etc. In this case, we have assumed that the residual effect is nil. But in some situation the residual effect is so prominent that the assumption is not valid. Following are the two methods by which we can eliminate the residual effect.

(a) A gap or rest period is maintained so that the effect due to treamtment will not be carried over to the next.

(b) The residual effect is eliminated by a special technique of analysis of variance. The first method is not practical because during the rest period we are to apply some control treatmentwhich may react with the earlier treatment. Also we may not have extra time for keeping gap period.

For the second method, the lay-out for the cross-over design with treatments, t may be even or odd number, may be described as suggested by William (1949) as below :

(i) For even number of treatments :

(a) The first column is written according to sequence 1, 2, t, 3, (t-1), 4........Thus for t=4, the first column is written as 1, 2, 4, 3, and for t = 6 the first column would be 1, 2, 6, 3, 5, 4. The numbers indicate the treatments.

(b) Next (t - 1) columns are obtained from the first column by successive addition of 1 but if the number exceeds t, t is to be subtracted from it.

For example; when t = 4, the 4 columns can be written as :

1	2	3	4
2	3	4	1
4	1	2	3
3	4	1	2

(ii) For odd number of treatments : In this case there will be two squares. The first column of one square is 1, 2, t, 3, (t - 1).....and the first column of the second square is the first column of the first square but in reverse order.

Thus for n = 5, two squares are as follows :

1st square

1	2	3	4	5
2	3	4	5	1
5	1	2	3	4
3	4	5	1	2
4	5	1	2	3

2nd square

4	5	1	2	3
3	4	5	1	2
5	1	2	3	4
2	3	4	5	1
1	2	3	4	5

Example 11.8 Three feeds A, B and C were given to six cows in three lactation stages. The plan and milk production in kg/day are given below. Test the effect of feeds on milk production. (Assuming that there is no residual effect).

	Cow 1	Cow 2	Cow 3	Cow 4	Cow 5	Cow 6
Stage-1	A - 10	C - 16	A - 12	B - 11	B - 14	C - 10
Stage - 2	C - 9	B - 7	B - 11	C - 10	A - 12	A - 13
Stage - 3	B - 14	A - 12	C - 10	A - 12	C - 8	B - 11.

Solution :

	Totals		Totals
Stage - 1	73	Cow - 1	33
Stage - 2	62	Cow - 2	35
Stage - 3	67	Cow - 3	33
Feed - A	71	Cow - 4	33
Feed - B	68	Cow - 5	34
Feed - C	63	Cow - 6	34

Correction factor (C. F.) $= \frac{202^2}{18} = 2266.89.$

Total S.S. $= 10^2 + \ldots\ldots + 11^2 -$ C. F. $= 2350 - 2266.89 = 83.11.$

S. S. due to Stage $= \frac{73^2 + 62^2 + 67^2}{6} -$ C. F. $= 2277 - 2266.89 = 10.11$

S.S. due to Feed $= \frac{71^2 + 68^2 + 63^2}{6} -$ C. F. $= \frac{13634}{6} -$ C. F. $= 2272.33 -$ C. F. $= 5.44.$

S. S. due to Cow $= \frac{33^2 + \ldots + 34^2}{3} -$ C. F. $= \frac{6804}{3} -$ C. F.$= 2268 -$ C. F. $= 1.11.$

Error S.S. $= 83.11 - 10.11 - 5.44 - 1.11 = 66.45.$

H_0 : The effects of all the feeds are same.

Table-11.22

ANOVA TABLE

Source of variation	d.f	S.S.	M.S.	F	5%F
Stage	2	10.11	5.06		
Feed	2	5.44	2.72	0.33	8.65
Cow	5	1.11	0.22		
Error	8	66.45	8.31		
Total	17	83.11			

The calculated value of F with (2,8) d.f. is smaller than the tabulated value of F at 5% level of significance. Therefore, the calculated value is insignificant and the hypothesis may be accepted.

11.4 Multiple Comparison Tests

The significant value of F for treatments, indicates the rejection of null hypothesis H_0 : The treatment effects are equal. In that case we may be interested in making comparison between pairs of treatment means and finally to decide the most effective one.

For the above purpose, following are the usual comparison tests.

Least Significant Difference (l. s. d) Test : It is the oldest method for making comparison of treatment means to see whether the difference of the observed means of treatment pairs exceeds the l.s.d. numerically. We declare the means of pair of treatments to be significantly different if the difference of treatment means exceeds l.s.d. which is calculated by $t_\alpha \times s\sqrt{\frac{2}{r}}$

where t_α is the value of Student's t with error d.f. at $100\alpha\%$ level of significance. s^2 is the M. S. of error and r is the number of replications of the treatments. For unequal replications ,r_1 and r_2,

$$\text{l. s. d} = t_\alpha \times s\sqrt{\frac{1}{r_1}+\frac{1}{r_2}}.$$

The test criterion is very easy to calculate but restricted in the sense that treatment pairs should be independent and are to be pre-determined. Therefore, it cannot be used for all possible pairs of treatment means.

Example 11.9 Apply l.s.d. test for testing the difference of treatment means of F_1 and F_6 from the data given in Example 11.2

Solution : We have,

$$\text{l.s.d.} = t_{0.01} \times \sqrt{\frac{2s^2}{r}}. \quad \text{where, } t_{0.01} = 2.947,\ s^2 = 0.469\ ;\ r = 4.$$

Mean of $F_1 = 5.18$ and $F_6 = 7.22$. Therefore, $7.22 - 5.18 = 2.04$.

$$\text{Now, l.s.d.} = 2.947 \times \sqrt{\frac{2 \times 0.469}{4}} = 1.427.$$

Therefore, the difference between the two means of F_1 and F_6 is highly significant indicating that F_6 is better than F_1.

Tukey's ω-Test : For comparing all possible pairs of treatment means we arrange the treatment means in ascending order of magnitude as $\bar{x}_1, \bar{x}_2, \ldots\ldots\ldots\ldots \bar{x}_t$. The studentized range statistic is given by,

$$q_{t,p} = \frac{(\overline{x}_t - \overline{x}_1)\sqrt{r}}{s}$$

where s^2 is the M.S. of error with p d. f. and r is the number of replications of all the treatments. The values of $q_{t,p}$ are available in Biometrika Tables, Vol.-1, Table-29.

For comprison all possible pairs of treatment means Tukey (1953) suggested the statistic, $\omega = q_{(\alpha)\,t,p} \times \frac{s}{\sqrt{r}}$.

For unequal replications, $\omega = q_{(\alpha)t,p} \times s\left[\frac{1}{2}\left(\frac{1}{r_1}+\frac{1}{r_2}\right)\right]^{\frac{1}{2}}$

where $q_{(\alpha)\,t,p}$ is the value of $q_{t,p}$ at upper $100\alpha\%$ point. $100\alpha\%$ level of significance generally depends on the original ANOVA table. Tukey's ω-test is very important since only one value like l.s.d. is used to compare all possible pairs of treatment means. ω is sometimes called honestly significant difference (h. s. d.) test.

Example 11.10 Apply Tukey's ω- test for testing all possible pairs of means for significance using the data given in Example 11.2.

Solution : We have, $\omega = q_{(\alpha)\,t,p} \times \frac{s}{\sqrt{r}}$.

where $\alpha = 0.01$, $t = 6$, $p = 15$, $r = 4$, $s^2 = 0.469$.

and $q_{(0.01)\,6,\,15} = 5.80$ Vide Biometrika Tables. Vol.-1, Table—29.

$$\therefore\ \omega = 5.80 \times \sqrt{\frac{0.469}{4}} = 1.986.$$

The treatment means corresponding to different treatments arranged in order of magnitude are :

F_1	F_2	F_5	F_3	F_4	F_6
5.178	5.195	5.795	6.180	7.013	7.223.

Treatments underscored by a common line donot differ significantly while the others differ significantly. Thus F_6 is significantly better than F_1 and F_2 at 1% level of significance. And there is no significant difference among F_3, F_4, F_5 and F_6.

Newman-Kewls' Sequential Range Test : In Tukey's ω-test the number of ordered steps between the means of the treatment are not considered. Considering this aspect, Newman-Kewls' put forward the following method by which the most effective treatment can be determined.

(i) Arrange the treatment means in ascending order of magnitude in a two-way table as below :

Table-11.23

	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_{t-1}$	$\bar{x}_t$
$\bar{x}_t$	$\bar{x}_t - \bar{x}_1$	$\bar{x}_t - \bar{x}_2$		$\bar{x}_t - \bar{x}_t$	0
$\bar{x}_{t-1}$	$\bar{x}_{t-1} - \bar{x}_1$	$\bar{x}_{t-1} - \bar{x}_2$		0	
|	|				
|	|				
|	|				
|	|				
$\bar{x}_2$	$\bar{x}_2 - \bar{x}_1$	0			
$\bar{x}_1$	0				

(ii) (a) The range $\bar{x}_t - \bar{x}_1$ is compared with critical difference

$$q(\alpha)\ (t=t-1+1,p)\frac{s}{\sqrt{r}}.$$

(b) If the test under (a) is significant, the difference to the right $(\bar{x}_t - \bar{x}_2)$ is compared with $q_{(\alpha)\ (t-1=t-2+1,p)}\frac{s}{\sqrt{r}}$.

(c) If the test under (b) is significant, the difference to the further right $(\bar{x}_t - \bar{x}_3)$ is compares with $q_{(\alpha)(t-2=t-3+1,p)}\frac{s}{\sqrt{r}}$.

(d) If the test under (c) is in significant we turn to the second row and proceed as in the first row and admit only up to the column where we get significant differences.

xample 11.11 Apply N. K. sequential range test for testing all possible pairs of means for significance, using the data given in Example 11.2.

Solution : We prepare the following table for calculating different $q_{(\alpha)t,p}\frac{s}{\sqrt{r}}$ for different values of t.

Table-11.24.

Value of t	$q_{(.01)\,t,15}$	$q_{(0.01)t,15} \times \frac{s}{\sqrt{r}}$
6	5.80	1.986
5	5.56	1.904
4	5.25	1.798
3	4.83	1.654
2	4.17	1.429.

Two-way table of differences of treatment means corresponding to different treatments are as follows :

Table-11.25

	F_6 7.223	F_4 7.013	F_3 6.180	F_5 5.795	F_2 5.195	F_1 5.178
5.178	2.045**	1.835	1.002	0.617	0.017	—
5.195	2.028**	1.818	0.985	0.600	—	—
5.795	1.428	1.218	0.205	—	—	—
6.180	1.043	0.833	—	—	—	—
7.013	0.210	—	—	—	—	—
7.223	—	—	—	—	—	—

The difference 2.045 is compared with 1.986, the difference 1.835 is compared with 1.904 and so on. Thus it is seen that F_6 is significantly better than F_1 and F_2.

Duncan's New Multiple Range Test : In the Newman-Kewls' (N,K) sequential range test we have considered a constant level of significance irrespective of the number of steps of the means are apart. Duncan (1955) made α_k, the level of significance a variable from test to test by considering the level of significance as $\alpha_k = 1 - (1 - \alpha)^{k-1}$ where k is the number of order steps between the ordered means and α is as defined earlier. We define $q_{(\alpha_k)}$ as the significant studentised range (S.S.R.) The value of S.S.R. is given in Duncan (1955). The least significant range (L.S.R) is defined by,

$$\text{L.S.R.} = \text{S.S.R} \times \frac{s}{\sqrt{r}}.$$

In case, a pair of means differs by more than its L.S.R, they are declared to be significantly different.

Example 11.12 Apply Duncan's new multiple range test of testing all possible pairs of means for significance using data given in Example 11.2.

Solution : The values of S.S.R. and L.S.R. for different values of k are as follows :

Table-11.26

Value of k	2	3	4	5	6
S.S.R.	4.17	4.37	4.50	4.58	4.64
L.S.R.	1.428	1.496	1.541	1.580	1.558

Treatment means corresponding to different treatments are arranged as follows :

F_1	F_2	F_5	F_3	F_4	F_6
5.178	5.195	5.795	6.180	7.013	7.223.

which indicates that F_6 and F_4 is significantly better than F_1 and F_2. The difference between any pair of underscored treatment means being insignificant.

11.5 Factorial Experiment

A certain character under study may be influenced by a number of factors at different levels and hence it is necessary to test different combinations of the levels of the factors. An experiment in which a number of factors at different levels are tested for their effects and interactions is called factorial experiment. There are two types of factorial experiment, symmetrical and asymmetrical.

Factorial experiments provides study not only the individual effects of each factor but also their interactions. In these experiments we require less resources to get same precision for each factor effect. They give an exploratory work and hence they are widely used in research work. They also form the basis of other designs of considerable practical importance.

When the number of factors are large in number, it is difficult to handle because, blocks of required size may not be available. In that case we can deal with fractional factorial. For this aspect, the serious readers may be referred to Montgomery (1976) and Jhon (1971).

Symmetrical Factorial : When the factors, each have the same number of levels, they are called symmetrical factorial experiment. For an example, let $F_1, F_2, \ldots\ldots F_n$ be n factors each at s levels, then we have a symmetrical factorial experiment of the type s^n.

2^n Factorial Experiment : Let us consider the factorial design of the type 2^n which has n factors each at 2 levels. For more simplicity sake, we consider $n = 2$ i.e, the most simple factorial experiment of the type 2^2. Let the two factors be denoted by A and B each at 2-levels, the low level is denoted by 0 and the high level is denoted by 1. The treatment combination 00 represents both the factors at the low level and may be denoted by (1), 10 represents A at high level and B at low level, may be denoted by a, 01 represents A at low and B at high level, may be denoted by b and 11 represents both the factors at high level, may be denoted by ab. Let us consider r replications. Further let the lower case letters (1), a, b and ab represent the total of the observations in all the r replicates corresponding to different treatment combinations.

Main-effect and Interaction-effect : When two factors A and B are involved in the experiment, the effect of A at the low level of B is $\frac{[a - (1)]}{r}$ and the effect of A at the high level of B is $\frac{[ab - b]}{r}$. Averaging both the quantities we have the main effect of A, denoted by, $A = \frac{1}{2r}[(ab - b) + (a - (1)].$

$$= \frac{1}{2r}[ab + a - b - (1)] = \frac{1}{2r}(a - 1)(b + 1)$$

Similarly the main effect of B is

$$B = \frac{1}{2r}[ab + b - a - (1)] = \frac{1}{2r}(a + 1)(b - 1).$$

Now the interaction AB is the average difference between the effect of A at the high level of B and the effect of A at the low level of B. Thus,

$$AB = \frac{1}{2r}[\{ab - b\} - \{a - (1)\}] = \frac{1}{2r}[ab + (1) - a - b] = \frac{1}{2r}(a - 1)(b - 1).$$

The interaction effect BA is seen to give the same expression as above and hence, Interaction AB = Interaction BA.

It is seen that the effects are expressed in term of contrasts of treatment combinations. As the three contrasts are mutually orthogonal, we can split the treatment sum of square with 3 d.f. into three sum of squares each with 1 d.f. corresponding to three effects. The contrasts representing the effects A, B and AB are shown below with + and - signs against the treatment combinations.

Table-11.27

Treatment Combinations		Factorial effects		
		A	B	AB
(1)	00	—	—	+
a	10	+	—	—
b	01	—	+	—
ab	11	+	+	+

2^3 Factorial Experiment : Let us consider three factors A, B and C each at 2 levels, designated as earlier. The treatment combinations can be written as (1), a,b, ab, c, ac, bc, and abc. In terms of 0 and 1 the treatment combinations can be written as 000, 100, 010 110, 001, 101, 011 and 111. As earlier the lower case letters indicate the total of observations corresponding to that particular treatment combination in r replications.

Main-effects and Interaction-effects : The effect of A when B and C are at low level is $\frac{[a - (1)]}{r}$, the effect of A when B is at high level and C is at low level is $\frac{[ab - b]}{r}$, the effect of A when B is at low level and C is at high level is $\frac{[ac - c]}{r}$ and finally the effect of A when both B and C are at high level is $\frac{[abc - bc]}{r}$. Thus the main effect of A is the average of these four effects

which is $A = \frac{1}{4r}[a - (1) + ab - b + ac - c + abc - bc]$

$= \frac{1}{4r}[a + ab + ac + abc - (1) - b - c - bc] = \frac{1}{4r}(a - 1)(b + 1)(c + 1).$

Similarly the main effect of B and C are as follows :

$B = \frac{1}{4r}[b + ab + bc + abc - a - c - ac - (1)] = \frac{1}{4r}(a + 1)(b - 1)(c + 1)$

and $C = \frac{1}{4r}[c + ac + bc + abc - a - b - ab - (1)] = \frac{1}{4r}(a + 1)(b + 1)(c - 1)$

When C is at low level, the interaction effect AB is the average difference in the effect of A at two levels of B i.e. $\frac{1}{2r}[(ab - b) - \{a - (1)\}]$

When C is at high level the interaction effect AB is $\frac{1}{2r}[(abc - bc) - (ac - c)]$

In case of three factors, A, B and C, the interaction effect of AB is therefore, the average of these two effects.

Thus $AB = \frac{1}{4r}[ab - b - a + (1) + abc - bc - ac + c]$

$= \frac{1}{4r}[abc + ab + c + (1) - a - b - ac - bc] = \frac{1}{4r}(a + 1)(b - 1)(c + 1)$, similarly

$AC = \frac{1}{4r}[abc + ac + b + (1) - a - c - ab - bc] = \frac{1}{4r}(a - 1)(b + 1)(c - 1)$

and $BC = \frac{1}{4r}[abc + bc + a + (1) - b - c - ab - ac] = \frac{1}{4r}(a + 1)(b - 1)(c - 1)$

AB, AC and BC are usually called 2-factor interaction effects. The interaction effect ABC is the average difference between AB interaction for two different levels of C.

Thus, $ABC = \frac{1}{4r}[\{(abc - bc) - (ac - c)\} - \{ab - b - a - (1)\}]$

$= \frac{1}{4r}[abc + a + b + c - ab - ac - bc - (1)] = \frac{1}{4r}(a - 1)(b - 1)(c - 1).$

ABC is called the 3-factor interaction. It can be shown that,
Int. ABC = Int. BCA = Int. ACB. Therefore, the order of the letters are immeterial in case of having interaction effects.

All main effects and interaction effects are expressed in terms of contrasts of treatment combinations and the contrasts are mutually orthogonal. The sum of squares due to treatments with 7 d.f. can be split up into different sum of squares each with 1 d.f. due to different effect components. The contrasts representing main effects A, B and C, 2-factor interaction effects AB, AC and BC and 3-factor interaction effect ABC are shown in Table-11.28 with + and - signs against the treatment combinations.

Table-11.28

Treatment Combinations	Factorial effects						
	A	B	C	AB	AC	BC	ABC
(1)	—	—	—	+	+	+	—
a	+	—	—	—	—	+	+
b	—	+	—	—	+	—	+
ab	+	+	—	+	—	—	—
c	—	—	+	+	—	—	+
ac	+	—	+	—	+	—	—
bc	—	+	+	—	—	+	—
abc	+	+	+	+	+	+	+

The Table-11.28 has several interesting properties :

(1) Every column has an equal number of + and — signs.

(2) The sum of products of co-efficient of signs in any two columns is zero.

(3) The product of signs of any two columns yields a column in the table.

For example, $A \times B = AB$ and $AB \times B = AB^2 = A$. We see that the products are formed modulus 2 (the exponent can only be zero or one if it is greater than one, it is reduced by multiples of two until it is either zero or one).

In general the main effects and interaction effects of 2^n factorial experiments can be obtained in the above way. The sum of squares of any effect is equal to

$\frac{(\text{Contrast})^2}{2^n \times r}$, where n is the number of factors and r is the number of replications. Thus getting the sum of squares of different components, the ANOVA table can be prepared for any experiment of 2^n series when it is conducted in any one of the basic designs.

Yate's Algorithm for the 2^n Factorial Experiments :

There is another systemetic method of getting the estimate of effects and the sum of squares of different effects usually known as Yates' Algorithm.

The procedures are as follows :

1. The treatment combinations are written as usual in a column.

2. The total of the responses (yields, measures of observations etc.) are written columnwise corresponding to each treatment combination.

3. The first half of the next column which is denoted by col-1 is obtained by adding the responses in adjacent pairs. The second half of col-1 is obtained by taking second value minus the first value in each pair.

4. Col-2 can be obtained from col-1 just as col-1 is obtained from response column.

The process of pairwise addition and subtraction is continued to get col-n if it is a 2^n factorial experiment.

5. The estimates of the the effects can be obtained dividing the values (avoiding the first one) of col-n corresponding to treatment combinations obtained by mentioned above procedure by $r \times 2^{n-1}$ where n and r are described earlier.

6. Sum of square of the effects can be obtained by squaring the value (avoiding the first one) of col-n corresponding to treatment combinations and dividing by $r \times 2^n$.

Thus the sum of squares of different effects are obtained. The replications and error sum of squares can be obtained as usual and the analysis of data in a 2^n factorial experiment conducted in any one of the basic designs in r replications can be performed.

Example 11.13 For a factorial experiment with three factors N, P and K each at two levels conducted in a randomised block design in 4 replications, the lay-out and yield per plot are given below :

Rep—1					Rep—2			
(1) 25	k 32	pk 24	p 27		p 32	nk 34	npk 42	(1) 44
nk 32	np 30	n 30	npk 36		n 46	np 30	k 39	pk 36

Rep—3					Rep—4			
k 32	pk 20	n 28	nk 28		np 32	nk 41	npk 45	k 35
npk 30	(1) 24	p 26	np 36		(1) 34	pk 39	n 41	p 29

Analyse the data and give your comment.

Solution : Grand Total of the observations=1059

$$\text{Correction Factor (CF)} = \frac{1059^2}{32} = 35046.2833.$$

Total S.S. = 36381 - CF = 1334.7167.

Replication S.S. = 35662.1250 - CF = 615.842.

S.S. due to different main-effects and interaction effects can be obtained from the following Yates' Algorithm :

Table-11.29

Treatment Combination	Total from all replicates	Col-1	Col-2	Col-3	Mean effect $=\frac{\text{Col-3}}{4 \times 2^2}$	S.S. $=\frac{[\text{Col-3}]^2}{4 \times 2^3}$
(1)	127	272	514	1059	—	—
n	145	242	545	63	3.9375	124.0312
p	114	273	32	-31	-1.9375	30.0312
np	128	272	31	33	2.0625	34.0312
k	138	18	-30	31	1.9375	30.0312
nk	135	14	-1	-1	-.0625	0.0312
pk	119	-3	-4	29	1.8125	26.2812
npk	153	34	37	41	2.5625	52.5313

H_0 : Effects of all the main effects are same and interaction effects are nil.

Table-11.30

ANOVA TABLE.

Source of variation	d.f.	S.S.	M.S.	F	5%F
Replication	3	615.842	205.2807	10.2176	
N	1	124.0312	124.0312	6.1735	4.32
P	1	30.0312	30.0312	1.4948	
K	1	30.0312	30.0312	1.4948	
NP	1	34.0312	34.0312	1.6939	
NK	1	0.0312	0.0312	0.0016	
PK	1	26,2812	26.2812	1.3081	
NPK	1	52.5313	52.5313	2.6147.	
Error	21	421.9062	20.0908		
Total	31	1334.7167			

Conclusion : The effect of nitrogen is seen to be significant and all other effects are insignificant.

3^n Factorial Experiments : When factors each have three levels instead of two, the scope of the experiment increases. It gives more information than the earlier because it provides the opportunity to study linear as well as quadratic effects. But it should be remembered that the treatment combinations increases rapidly as the number of levels per factor increases.

Here we are considering n factors each at 3 levels. For simplicity sake, let us consider n = 2 i. e. 2 factors each at 3 levels giving a 3^2 factorial experiment. Let the two factors be denoted by A and B and 3 levels be coded by 0, 1 and 2. The treatment combinations can be written in two different ways namely :

(1), a_1, a_2, b_1, a_1b_1, a_2b_1, b_2, a_1b_2 and a_2b_2 and 00, 10, 20, 01, 11, 21, 02, 12 and 22.

These treatment combinations can be alloted at random to plots in any one of the designs. The main effects and interaction effects can be expressed in the method given below :

Considering a single factor A, $(a_1 - a_0)$ indicates the response at the level 0 and that of $(a_2 - a_1)$ at the level 1. The sum of these two responses gives the linear effect $(a_2 - a_0)$ and their difference gives the quadratic effect $(a_2 - 2a_1 + a_0)$. Thus linear and quadratic effects of B can also be defined. The interaction effect can be split into components of interactions between linear and quadratic effects of the two factors. Denoting the linear and quadratic effects of A by A_l and A_q respectively and similarly for B the four interaction components each with 1 d.f. can be written as, (without the divisors),

$$A_lB_l = (a_2 - a_0)(b_2 - b_0)$$

$$A_lB_q = (a_2 - a_0)(b_2 - 2b_1 + b_0)$$

$$A_qB_l = (a_2 - 2a_1 + a_0)(b_2 - b_0)$$

$$A_qB_q = (a_2 - 2a_1 + a_0)(b_2 - 2b_1 + b_0)$$

Thus it is seen that the main effects and interaction effects can be expressed in terms of contrasts which are mutually orthogonal and therefore the treatment sum of squares for different components can be obtained from the sum of squares due to treatments.

Yates' Algorithm for the 3^n Factorial Experiments : The estimates and sum of squares of different components of effect in 3^n factorial experiment can be obtained by Yates' Algorithm as follows :

(1) The treatment combinations are written in the systematic manner in a column.

(2) The total of the responses are written column wise corresponding to each of treatment combination. The first one third of the next column denoted by col-1 consists of the sum of each of the sets of three values in the response column. The second one third of Col-1 is obtained by the third value minus the first value in the sets of three values. This operation computes the linear components of the effects. The last one third of the column is obtained by taking the sum of first and third values minus twice the second value in each set of three values. This computes the quadratic components.

(3) The process is to be carried out n times to give Col-n giving the estimates of effects in 3^n factorial experiment without considering the divisors.

The devisors for sum of squares for different treatment effects are obtained from $2^p 3^q r$ where p is the number of factors in the effect considered and q is the number of factors in the experiment minus the number of linear terms in this effect and r is the number of replications.

In this way, the sum of squares of different effects are computed, the replication sum of squares and error sum of squares can be computed as usual and the analysis of the 3^n factorial experiment can be performed.

Confounding : We usually recommend that the factorial experiments can be conducted in any one of the basic designs. We have seen that the data in these experiments are analysed by spliting the treatment components in main effects and interaction effects.

When the number of factors and/or the number of levels of a factor increases, it becomes almost difficult to conduct the experiment with suitable size of the blocks. In this case, the contrast of the treatment combinations of some interactions effects usually of higher order interactions are divided into some parts and the treatment combinations are alloted at random to seperate blocks giving a replication and thus the size of the blocks are reduced to managable number. In such cases, contrasts of the interactions and contrasts between the block totals give the same function. The contrasts are therefore mixed up with the block effects and can not be separated. In other words, the interactions effects have been confounded with blocks. This device of reducing the block size by making one or more interaction contrasts identical with block contrasts, is known as confounding.

Total and Partial Confounding : When there are two or more replications, a question arises whether the same interaction is confounded in each replication or different sets of interactions are confounded in different replications. Both the procedures are practiced. If the same set of interactions is confounded in all the replications, confounding is called total. In such confounded factorial experiment, the estimate of interaction effects confounded, cannot be obtained but all other main effects and interaction effects can be estimated with better precision because of reduced block size. If again different sets of interactions are confounded in different replications, confounding is called partial. In such method of confounding the informations of the confounded interaction effects can be recovered from those replications in which they are not confounded.

Let us consider an example each from 2^n and 3^n series of factorial experiments.

(i) Let us consider 2^3 factorial experiment in which the factors are represented by A, B and C each at 2 levels. One way of writing the treatment combinations are (1), a, b, ab, c, ac, bc, abc. When the highest order interaction effect, ABC is confounded, the two block contents can be obtained by choosing even number of letters common with the effect and the other by choosing the odd number of letters common with the effect. Therefore the block contents can be written as,

Bl-1	Bl-2
(1)	a
ab	b
ac	c
bc	abc

If we consider the levels by 0 and 1, the treatment combinations can be written as 000, 100, 010, 110, 001, 101, 011, 111.

Again considering the effect ABC to be confounded, two block contents can be obtained by solving two equations respectively

$$x_1 + x_2 + x_3 \left.\begin{array}{l} = 0 \\ = 1 \end{array}\right\} \bmod 2;$$

Bl-1	Bl-2
000	100
011	010
110	001
101	111

(ii) Let us consider 3^3 factorial experiment. Here we consider three factors, A, B, and C each at 3 level, denoted by 0, 1 and 2, the treatment combinations can be written as 000, 100, 200, 010, 110, 210, 020, 120, 001, 101, 201, 011, 111, 211, 021, 121, 002, 102, 202, 012, 112, 212, 022, 122, 220, 221, 222.

Let the interaction effect AB_2C be confounded with blocks. In this case we get 3 blocks in a replication and these can be obtained by solving three equations namely,

$$x_1 + 2x_2 + x_3 \left.\begin{aligned} &= 0 \\ &= 1 \\ &= 2 \end{aligned}\right\} \text{mod } 3$$

The block contents are as follows :

Block—1	Block—2	Block—3
000	001	002
011	012	010
022	020	021
110	111	112
102	100	101
121	122	120
201	202	200
212	210	211
220	221	222

The block containing 000 is generally called principal block. Once it is obtained, the second block can be obtained by adding 1 mod 3 to the last element of the first block contents and the third can be obtained by adding 2 mod 3 to the last element of the first block contents or by adding 1 mod 3 to the last element of the second block contents.

If AB_2C is confunded in 3 replications, say, the effect AB_2C is totally confounded and the information due to AB_2C is completely lost. But if AB_2C is confounded in the first replication. ABC_2 is confounded in the second replication and AB_2C_2 is confounded in the third replication then neither of the effects is totally confounded as the estimate of AB_2C can be obtained from the second and third replications, the estimate of ABC_2 can be obtained from the first and third replications and lastly the estimate of AB_2C_2 can be obtained from the first and second replications. Hence in this case, the effects namely AB_2C, ABC_2 and AB_2C_2 are partially confounded.

Confounding more than one effect : With the increase of the number of factors, the treatment conbinations increase sharply. In that case, 2 blocks in case of 2^n series and 3 blocks in case of 3^n series may not surve our purpose of getting blocks of suitable size. That is, if we are to reduce the size of the blocks more than that obtained earlier, we are to confounded more than one higher-order interaction effects.

For 2^n series, when we are to get 2^k blocks of size 2^{n-k} in a replication, $2^k - 1$ interaction effects are to be confounded of which k effects are independent and the remaining (2^k - k - 1) are generalised effects. For example, in a 2^5 factorial experiment if we are to get 2^2 block of size 2^3 in a replication, 2 interaction effects are to be confounded, say ABC and BCDE, the number of generalised effect is (2^2 - 2 - 1) i.e. 1 which is ABC x BCDE = AB^2C^2DE = ADE.

The block contents can be obtained by solving the following two sets of equations simultaneously.

$$\left.\begin{aligned} x_1 + x_2 + x_3 &= 0 \\ &= 1 \end{aligned}\right\} \bmod 2$$

$$\left.\begin{aligned} x_2 + x_3 + x_4 + x_5 &= 0 \\ &= 1 \end{aligned}\right\} \bmod 2$$

i.e. the block contents of 4 blocks can be obtained from the solutions of the following equations in terms of treatment combinations :

$$\left.\begin{aligned} x_1 + x_2 + x_3 &= 0 \\ x_2 + x_3 + x_4 + x_5 &= 0 \end{aligned}\right\} \bmod 2$$

$$\left.\begin{aligned} x_1 + x_2 + x_3 &= 0 \\ x_2 + x_3 + x_4 + x_5 &= 1 \end{aligned}\right\} \bmod 2$$

$$\left.\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + x_3 + x_4 + x_5 &= 0 \end{aligned}\right\} \bmod 2$$

$$\left.\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + x_3 + x_4 + x_5 &= 1 \end{aligned}\right\} \bmod 2$$

Similar explanation is given in detail for 3^n and in general s^n factorial experiment in Das and Giri (1979).

When the number of treatment combinations are large in number, fraction of the factorial experiment can be taken into consideration and experiments with blocks of small size can be handled. This type of design is called fractional factorial design which is beyond the scope of this text. Reference on this regard can be made from Das and Giri (1979) and Montgomery (1976).

The sum of squares of all the effects are obtained by 'Yates' Algorithm. The sum of squares of confounded effects will give us the block sum of squares. The degree of freedom for block is equal to the number of effects confounded. All other components can be obtained as usual. The sum of squares due to those affected interactions will be absent in the analysis of variance table. In case of analysis of partialy confounded factorial experiment, the sum of square of the effects which are not affected can be obtained by the usual Yates' Algorithm. The sum of squares of the affected effects can be obtained by Yates' Algorithm from the replications where the concerned effects are affected. In such case, there is another source of variation namely 'Blocks within replicates' whose sum of squares can be obtained from the addition of sum of squares of affected effects from the corresponding replicates where they are affected.

Example 11. 14 The plan and yield per plot (in a suitable unit) of 2^3 field experiments on wheat are given below : the treatments being all combinations of two levels of drug D (0,1), two levels of potash K(0,1) and two levels of superphosphate P(0,1). The experiment was conducted in four replications each having two blocks. Detect the effects confounded in different replicates and analyse the data.

Rep—1

Block-1	000 (1) 32	111 pkd 38	011 kd 26	100 p 35
Block-2	001 d 43	010 k 45	101 pd 45	110 kp 31

Rep—2

Block-3	101 pd 40	000 (1) 48	010 k 40	111 pkd 31
Block-4	110 pk 34	100 p 48	011 kd 31	001 d 33

Rep—3

Block-5	111 pkd 42	110 pk 44	001 d 32	000 (1) 46
Block-6	010 k 33	101 pd 47	100 p 42	011 kd 29

Rep—4

Block-7	000 (1) 37	110 pk 24	011 kd 32	101 pd 40
Blocks-8	100 p 42	111 pkd 42	010 k 29	001 d 34

Solution : From the block contents, it is seen that KD is confounded in Replicate-1, PD is confounded in Replicate-2. PK is confounded in Replicate-3 and lastly PKD is confounded in Replicate-4.

Grand Total=1195. Correction factor (C. F.) $= \frac{1195^2}{32} = 44625.781$.

The block totals are : $B_1 = 131$, $B_2 = 164$, $B_3 = 159$, $B_4 = 146$,

$B_5 = 164$. $B_6 = 151$, $B_7 = 133$ and $B_8 = 147$.

Total S.S. = 46041 - C. F. = 1415.219,

Block S.S. = $\frac{(131)^2 + \ldots\ldots ++ (147)^2}{4}$ - C. F. = 44912.25 - C. F. = 286.469.

To obtain the treatment S. S. we prepare the following table :

Table-11.31

(1) Treatment combina-tion	(2) Total from all replicates	(3) Total from replicates 1, 2 and 3	(4) Total from replicates 1, 2 and 4.	(5) Total from replicates 1, 3, and 4,	(6) Total from replicates 2, 3, and 4.
(1)	163	126	117	115	131
p	167	125	125	119	132
k	147	118	114	107	102
pk	133	109	89	99	102
d	142	108	110.	109	99
pd	172	132	125	132	127
kd	118	86	89	87	92
pkd	153	111	111	122	115

Main effects due to P, K and D (unaffected effects) can be obtained from column (2) of Table 11.31

[P] = - [1] + [p] - [k] + [pk] - [d] + [pd] - [kd] + [pkd] = 55.

[k] = - [1] - [p] + [k] + [pk] - [d] - [pd] + [kd] + [pkd] = - 93.

[D] = - [1] - [p] - [k] - [pk] + [d] + [pd] + [kd] + [pkd] = - 25.

The interaction effect PK is obtained from column (4) of Table 11.31

[PK] = [1] - [p] - [k] + [pk] + [d] - [pd] - [kd] + [pkd] = - 26.

The interaction effects PD is obtained from column (5) of Table 11.31

[PD] = [1] - [p] + [k] - [pk] - [d] + [pd] - [kd] + [pkd] = 62.

The interaction effect KD is obtained from Column (6) of Table 11.31

[KD] = [1] + [p] - [k] - [pk] - [d] - [pd] + [kd] + [pkd] = 40.

The interaction effect PKD is obtained from Column (3) of Table 11.31

[PKD] = - [1] + [p] + [k] - [pk] + [d] - [pd] - [kd] + [pkd] = 9.

All these effects are obtained without considering the divisors.

Now we compute sum of squares due to different main effects and affected interaction effects as usual.

$$\text{S.S. due to P} = \frac{[P]^2}{32} = 94.531.$$

$$\text{S.S. due to K} = \frac{[K]^2}{32} = 270.281.$$

$$\text{S.S. due to D} = \frac{[D]^2}{32} = 19.531.$$

$$\text{S.S. due to PK} = \frac{[PK]^2}{24} = 28.167.$$

$$\text{S.S. due to PD} = \frac{[PD]^2}{24} = 160.167.$$

$$\text{S.S. due to KD} = \frac{[KD]^2}{24} = 66.667.$$

$$\text{and S.S. due to PKD} = \frac{[PKD]^2}{24} = 3.375$$

H_0 : There are significant main effects and the interaction effects are nil.

Table-11.32

ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F.	5%F	1%F
Blocks	7	286.469	40.924	1.431		
P	1	94.531	94.531	3.306		
K	1	270.281	270.281	9.454		8.40
D	1	19.531	19.531	0.683		
PK	1	28.167	28.167	0.985		
KD	1	66.667	66.667	2.332		
PD	1	160.167	160.167	5.602	4.45	
PKD	1	3.375	3.375	0.118		
Error	17	486.031	28.590			
Total	31	1415.219				

From the above table it is seen that the main effect K is highly significant and the interaction effect PD is significant at 5% level of significance. All other main-effects are insignificant and other interaction effects are nil.

Analysis of 2^n Factorial Experiment in a Single Replication : When the number of factors is large the treatment combinations become very large. For example, a 2^5 has 32 treatment combinations, a 2^6 has 64 treatment combinations and so on. Since resources are usually limited, the number of replicates that the experimenter can employ may be restricted. Frequently available resources may only allow a single replicate of the design unless the experimenter is willing to omit some of the original factors.

With only a single replicate in a 2^n factorial experiment it is imposible to compute the mean square due to error. Thus it seems that hypothesis regarding main effects and interaction effects can not be tested. However, the usual approach to the analysis of a single replicate of the 2^n experiment is to assume that some of the higher order interactions to be negligible and the total of their sum of squares will give the estimate of sum of squares due to error. Thus the analysis can be performed. The degrees of freedom for error will be equal to the number of effects considered to be negligible.

But the practice of combining higher order interaction sum of squares should be done after proper varification becaue if some of these interactions are significant then the estimate of error will be inflated. Therefore, the experimenter must use both his knowledge of the phenomena under study and common sense in the analysis of such a design. A scientific method of detecting the insignificant effects was given by Daniel (1959). Assuming the data are normally and independently distributed, the $2^n - 1$ estimate of 2^n design are normally distributed. The method is to arrange the estimates of the effects in ascending order and plot the jth of these ordered values against $P_j = \frac{j - .5}{2^n - 1}$, j = 1, 2,........., $2^n - 1$, on normal probability paper. The effects, which are negligible, will tend to fall along a straight line on this graph, while significant effects will be far from the line. The negligible effects can thus be combined to form an estimate of error and the analysis of the data can be carried out.

Asymmetrical Factorial Experiment : When the number of levels of the factors are not same we get an asymmetrical factorial experiment. For example, the first factor, F_1 may have s_1 levels, second factor F_2 may have s_2 levels and so on the nth factor F_n may have s_n levels then the experiment of the type $s_1 \times s_2 \times \ldots \times s_n$ is called asymmetrical factorial experiment. Again if m factors each has s_1 level, n factors each has s_2 level, and so on, p factors each has s_k levels then the experiment denoted by $s_1^m \times s_2^n \times \ldots \times s_k^p$ is also called asymmetrical factorial experiment.

Symmetrical factorial experiment is some what inflexible because here all the factors have to be at the same number of levels. This may sometimes contradict the requirements of a practical experimenter. It may even be unrealistic in some situations to take all factors under investigation at the same number of levels. The above drawback can easily be overcome by adopting asymmetrical factorial experiment which is more flexible to meet the requirements of the experimenter.

Analysis of 3 x 2 Asymmetrical Factorial Experiment : 3 x 2 asymmetrical factorial experiment is the most simplest one, for which the procedure of analysis is given below :

Let there be two factors A at 3 levels and B at 2 levels. Denoting the levels of A by a_0, a_1 and a_2 and those of B by b_0 and b_1 the six treatment combinations of the factorial are $a_0 b_0$, a_0b_1, a_1b_0, a_1b_1, a_2b_0 and a_2b_1. These six treatment combinations can be accomodated in a block so that a randomised block design with r blocks can be constructed easily. The total degrees of freedom can be partitioned as follows :

Table-11.33

Source of variation	d.f.
Block	r - 1
Treatment	5
A	2
B	1
AB	2
Error	5(r - 1)
Total	6r - 1

The sum of squares of different components such as block, treatment and error can be obtained exactly in the same way as in the analysis of randomised block designs. The sum of squares due to main effects A and B and their interaction AB can be obtained by forming the following (A x B) table with six treatment totals, T_{ij}, i = 0, 1, 2, and j = 0, 1

Table-11.34

		Levels of A			
		a_0	a_1	a_2	Total
Levels of B	b_0	T_{00}	T_{10}	T_{20}	B_0
	b_1	T_{01}	T_{11}	T_{21}	B_1
	Total	A_0	A_1	A_2	G

Sum of squares due to A = $\dfrac{A_0^2 + A_1^2 + A_2^2}{2r}$ - C. F. where C.F. = $\dfrac{G^2}{6r}$

Sum of squares due to B = $\dfrac{B_0^2 + B_1^2}{3r}$ - C.F.

Sum of squares due to AB = Total S.S. due to (A x B) table

- S. S. due to A - SS due to B,

where Total S.S. due to (A x B) Table = $\dfrac{T_{00}^2 + T_{10}^2 + \ldots\ldots + T_{21}^2}{r}$ - C.F.

Thus the procedure of analysis of simple asymmetrical factorial design with number of treatment combinations those can be accomodated in a block is shown. For large number of treatment combinations, the procedure of confounding is also applicable here. Das and Giri (1979) can be referred on this regard.

11.6 Split-Plot Design

This is an special type of a symmetrical factorial design in which one factor requires bigger plots than the others for the convenniences of the experimenter. For example, if we have two factors namely irrigation and nitrogen fertiliser, it is convenient to apply irrigation to bigger plots and nitrogen to smaller plots, may be obtained by spliting the bigger plots into number equal to the levels of nitrogen fertiliser. Thus a replication is obtained with different sizes of experimental units for different treatments in the same experiment. We may have more than one replications and this type of design may be called split-plot design.

For this type of design first a randomised block design with bigger plots is taken to accomodate the factors which require bigger plots. Next each of the bigger plots is split into as many plots as the number of treatment coming from the other factor. The bigger plots are called main-plots and the treatments given to these are main-plot treatments or simply main treatments. The constituent parts of the main plots are called sub-plots and the treatments given to them are called sub-plot treatments. It is to be remembered that the different types of treatments are alloted at random to these respective plots. Therefore, split-plot design may be called the combination of two or more randomised block designs.

The analysis of the design is a bit complicated due to presence of two error components. The first error component is used to calculate F for main treatments and second error component is used to calculate F for sub-plot treatments and interaction effect of main plot and sub-plot treatments, thus giving an efficient test for latter case conducted, for that important treatments are confounded in the sub-plots. Due to the method of construction, the main treatments are usually confounded.

Analysis : Let there be p levels of main treatment A, q levels of sub-plot treatment B and there are r replications. Let y_{ijk} be the observation for the jth level of A, kth level of B and in the ith replication.

$i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, p$; and $k = 1, 2, \ldots, q$.

At the first step we prepare a two-way table like Main-treatment x Replication

from which the totals, $y_{i..} = \sum_{j}\sum_{k} y_{ijk}$; $y_{.j.} = \sum_{i}\sum_{k} y_{ijk}$, $y_{ij.} = \sum_{k} y_{ijk}$

and G = Grand total of all the observations can be obtained.

Now we calculate, Total S.S . $= \sum_{i}\sum_{j} y^2_{ijk}$ - C. F. where C. F. $= \dfrac{G^2}{pqr}$.

Replications S. S. $= \sum_{i} \dfrac{y_{i..}^2}{pq}$ - C.F. S. S. due to A $= \sum_{j} \dfrac{y_{.j.}^2}{rq}$ - C. F.

Error (1) S. S. $= \sum_{i}\sum_{j} \dfrac{y_{ij}^2}{q}$ - C. F. - Replications S. S. - S. S. due to A.

In the next step, we again prepare a two-way table like

Main-treatment x Sub-plot treatment.

The totals $y_{..k} = \sum_{i}\sum_{j} y_{ijk}$ and $y_{.jk} = \sum_{i} y_{ijk}$ etc. can be obtained.

Now we calculate, S. S. due to B $= \sum_{k} \dfrac{y_{..k}^2}{pr}$ - C. F.

S. S. due to AB $= \sum_{j}\sum_{k} \dfrac{y_{.jk}^2}{r}$ - C. F. - S. S. due to A - S. S. due to B.

Error (2) S. S. can be obtained as usual by subtraction.

The analysis of variance table can be furnished as given in Table-11.35.

In this case we test the hypothesis of equality of effects in sub-plot treatment and interaction effect to be nil.

Table-11.35

ANOVA TABLE

Source of variation	d.f.	Sum of squares
Replication	r - 1	$\frac{\Sigma y_{i..}^2}{pq} - \frac{G^2}{rpq}$
Main treatment	p - 1	$\frac{\Sigma y_{.j.}^2}{rq} - \frac{G^2}{rpq}$
Error (1)	(r - 1)(p - 1)	$\sum_i \sum_j \frac{y_{ij.}^2}{q} - \frac{\Sigma y_{i..}^2}{qp} - \frac{\Sigma y_{.j.}^2}{rq} + \frac{G^2}{rpq}$
Sub-plot treatment	q - 1	$\frac{\Sigma y_{..k}^2}{pr} - \frac{G^2}{rpq}$
Interaction (AB)	(p - 1)(q - 1)	$\sum_j \sum_k \frac{y_{.jk}^2}{r} - \frac{\Sigma y_{.j.}^2}{rq} - \frac{\Sigma y_{..k}^2}{pr} + \frac{G^2}{rpq}$
Error (2)	p(q - 1)(r - 1)	By subtraction
Total	pqr - 1	$\sum_i \sum_j y_{ijk}^2 - \frac{G^2}{rpq}$

Extension of the split - plot design : Split - plot design can be extended further by again spliting the sub-plots called second order sub-plots to assign at random to a further set of treatments. This type of design is called split-split-plot design. The analysis can be carried out in the same line as before with additional estimation of error component, called error (3) for the second - order sub-plots. This error (3) mean square is used for testing the effect of the second order sub-plot treatments and interactions with all other factors. The last mentioned effects would be estimated with the greatest precision as a result of the most efficient local control.

Example 11.15 In a varietal cum-manurial experiment on Soybeen, four levels of nitrogen 0, 0.1, 0.3 and 0.5 (kg) per plot, designated as n_0, n_1, n_2 and n_3 respectively were applied to each of three varieties V_1, V_2, V_3. The different levels of the manure for each variety were applied by spliting the plot into four sub-plots. The yields (in lbs) are given below in a systemetic pattern. Analyse the data.

Yield of the Split-plot experiment

	Rep-I		Rep-II		Rep-III		Rep-IV	
	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1
	104	105	117	129	123	123	105	135
V_1								
	n_2	n_3	n_2	n_3	n_2	n_3	n_2	n_3
	112	146	153	139	151	164	129	143
	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1
	112	109	111	123	117	109	124	129
V_2								
	n_2	n_3	n_2	n_3	n_2	n_3	n_2	n_3
	125	161	134	141	159	157	133	139
	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1
	116	119	119	132	102	116	135	143
V_3								
	n_2	n_3	n_2	n_3	n_2	n_3	n_2	n_3
	121	159	148	149	167	161	142	158

Solution : We know, C. F. = 838729.69

Total S.S. = 854631 - 838729.69 = 15901.31

Table-11.36

Main-Plot x Replication Table

	Rep I	Rep II	Rep III	Rep IV	Total
V_1	467	538	561	514	2080
V_2	507	509	542	525	2083
V_3	515	543	546	578	2182
Total	1489	1590	1649	1617	6345

S. S. due to Rep. $= \dfrac{1489^2 + \dots\dots + 1617^2}{12}$ - C. F. = 839925.92 - C. F. = 1196.23

S. S. due to Main-Plot treatment (variety) $= \dfrac{2080^2 + \dots + 218^2}{16}$ - C. F.

= 839150.81 - C. F. = 421.12.

Total S. S. from Rep. x Main-plot Table $= \dfrac{467^2 + \dots\dots + 578^{\ 2}}{4}$ - C. F.

= 841060.75 - C. F. = 2331.06.

(Rep. x Main-plot) Int. S.S. (E_1) = 2331.06 - 421.12 - 1196.23 = 713.71.

Table-11.37

Main-Plot x Sub-plot treatment Table

	V_1	V_2	V_3	Total
n_0	449	464	427	1385
n_1	494	470	510	1474
n_2	545	551	573	1669
n_3	592	598	627	1817
Total	2080	2083	2182	6345

Total S. S. from Main-plot x sub-plot treatment Table

$$= \frac{449^2 + ...+627^2}{4} - \text{C. F.} = 848717.25 - \text{C. F.} = 9987.56$$

$$\text{S. S. due to sub-plot treatment} = \frac{1385^2 + ... + 1817^2}{12} - \text{C. F.}$$

$= 848162.58 - \text{C. F.} = 9432.89$

Main-plot x Sub-plot treatment Int. S.S. = 9987.56 - 421.12 - 9432.89 = 133.55

Error (E_2) S.S. = 15901.31 - 1196.23 - 421.12 - 713.71 - 9432.89 - 133.55 = 4003.83

H_0: (i) Effects of all the four levels of nitrogen are equal.

(ii) There is no interaction effect between main-plot and sub-plot treatment.

Table-11.38

ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F	5%F	1%F
Replication (R)	3	1196.23				
Main-plot treat. (V)	2	421.12				
Int. (V x R) E_1	6	713.71				
Sub-plot (N)	3	9432.89	3144.293	21.2	—	4.60
Int. (NV)	6	133.55	22.258	0.15	2.46	—
Error (E_2)	27	4003.83	148.289			
Total	47	15901.33				

Since the calculated value of F with 3 and 27 d.f. corresponding to sub-plot treatment i.e. nitrogen is highly significant and therefore the hypothesis (i) may be rejected. But the calculated value of F corresponding to interaction between main-plot and sub-plot treatment is insignificant and hence the hypothesis (ii) may be accepted.

11.7 Strip-Plot Design

There are situations when both the factors require large plots with one set of plots superposed over the other sets at right angles, we get a strip-plot design. Let us consider an example having two factors like spacing and ploughing where the use of small plots by spliting bigger plots is not convenient. A block may be divided into strip in one direction to allocate one set of treatments called first factor, say, different spacing and into another set of strips in a direction at right angle to the first, to be alloted to the second set of treatments called second factor, say, ploughing. Any of the set of strips may further be divided into narrower strips for accomodating a new set of treatments called third factor. The allotment of the treatments to the strips are done at random at each stage. When we consider three factor, we get strip-strip-plot design.

Analysis : Like split-plot design we have to estimate error variance corresponding to each plot size in strip-plot design. In the above example, let three different plot sizes are involved ; different types of spacing constitute treatments, those have been alloted to plots of one size viz, the column strips, the ploughing treatments have been assigned to plots of second size namely, the row strips and lastly the comparisons of the different combinations of the two treatments or the interaction comparison have to be made from plots of third size formed by the interaction of the two sets of strips.

For the purpose of analysis of data in the above strip-plot experiment we have to prepare three two-way tables namely :
replication x first factor ; replication x second factor and first factor x second factor.

Let y_{ijk} be the observation for the jth level of first factor, kth level of second factor in the ith replication.

$i = 1, 2, \ldots\ldots r, ; j = 1, 2, \ldots\ldots p$ and $k = 1, 2, \ldots\ldots q$.

From the first two-way table replication x first factor, we get the following totals,

$$y_{i..} = \sum_{j}\sum_{k} y_{ijk} \; ; \; y_{.j.} = \sum_{i}\sum_{k} y_{ijk} \text{ and } y_{ij.} = \sum_{k} y_{ijk}$$

The correction factor (C. F.) = $\dfrac{G^2}{pqr}$ where G is grand total of all the observations i.e. $G = \sum_{i} y_{i..} = \sum_{j} y_{.j.}$

S. S. due to first factor = $\sum\limits_{j} \frac{y_{\cdot j \cdot}^2}{rq}$ - C. F.

Replications S. S. = $\frac{\sum y_{i \cdot\cdot}^2}{pq}$ - C. F.

Interaction effect between first factor and replication is considered as Error (1).

Thus Error (1) S.S. = $\sum\limits_{i}\sum\limits_{j} \frac{y_{ij \cdot}^2}{q}$ - C. F. - SS due to First factor - Replication S. S.

$= \sum\limits_{i}\sum\limits_{j} \frac{y_{ij \cdot}^2}{q} - \sum\limits_{j} \frac{y_{\cdot j \cdot}^2}{rq} - \sum\limits_{i} \frac{y_{i \cdot\cdot}^2}{pq} + \text{C. F.}$

Next, from second two-way table of Replication x Second factor we get the following totals.

$y_{\cdot\cdot k} = \sum\limits_{i}\sum\limits_{j} y_{ijk}$ and $y_{i \cdot k} = \sum\limits_{j} y_{ijk}$

S. S. due to second factor = $\sum\limits_{k} \frac{y_{\cdot\cdot k}^2}{rp}$ - C. F.

Interaction effect between second factor and replications is considered to be Error (2).

Error (2) S. S. = $\sum\limits_{i}\sum\limits_{k} \frac{y_{i \cdot k}}{p}$ - C. F. -S. S. due to the second factor

Replication S. S. = $\sum\limits_{i}\sum\limits_{k} \frac{y_{i \cdot k}^2}{p} - \sum\limits_{k} \frac{y_{\cdot\cdot k}^2}{rp} - \sum\limits_{i} \frac{y_{i \cdot\cdot}^2}{pq} + \text{C. F.}$

From the third two-way table of First factor x Second factor, we get the following new total $y_{\cdot jk} = \sum\limits_{i} y_{ijk}$

The interaction effect between First factor and Second factor can be computed as follows :

Interaction of First x Second factor S. S. = $\sum\limits_{j}\sum\limits_{k} \frac{y_{\cdot jk}^2}{r}$ - C. F. - S. S. due to First factor-S. S. due to Second factor $= \sum\limits_{j}\sum\limits_{k} \frac{y_{\cdot jk}^2}{r} - \sum\limits_{j} \frac{y_{\cdot j \cdot}^2}{rq} - \sum\limits_{k} \frac{y_{\cdot\cdot k}^2}{rp} + \text{C. F.}$

Total S. S. = $\sum\limits_{i}\sum\limits_{j}\sum\limits_{k} y_{ijk}^2$ - C. F.

Error (3) is obtained by usual subtraction.

The analysis of variance table can be furnished as given in Table-11.39 for testing hypothesis regarding the equality of effect of levels of First factor and Second factor and the interaction effects to be nil.

Table-11.39
ANOVA TABLE

Source of variation	d.f	Sum of squares
Replication	(r - 1)	$\sum_i \frac{\sum y_{i..}^2}{pq} - \frac{G^2}{rpq}$
First factor	(p - 1)	$\sum_j \frac{\sum y_{.j.}^2}{rq} - \frac{G^2}{rpq}$
Error (1)	(r - 1)(p - 1)	$\sum_i\sum_j \frac{y_{ij.}^2}{q} - \frac{\sum y_{i..}^2}{pq} - \frac{\sum y_{.j.}^2}{rq} + \frac{G^2}{rqp}$
Second factor	(q - 1)	$\sum_k \frac{\sum y_{..k}^2}{pr} - \frac{G^2}{rpq}$
Error (2)	(r - 1)(q - 1)	$\sum_i\sum_k \frac{y_{i.k}^2}{p} - \sum_i \frac{y_{i..}^2}{pq} - \sum_k \frac{y_{..k}^2}{rp} + \frac{G^2}{rpq}$
Interaction First x Second factor	(p - 1)(q - 1)	$\sum_j\sum_k \frac{y_{.jk}^2}{r} - \frac{\sum y_{.j.}^2}{rq} - \frac{\sum y_{.k}^2}{pr} + \frac{G^2}{rpq}$
Error (3)	(r - 1) (p - 1)(q - 1)	By subtraction
Total	rpq - 1	$\sum_i\sum_j\sum_k y_{ijk}^2 - \frac{G^2}{rpq}$

Hints of extension of strip-plot design is given earlier and the procedure of analysis can be carried out in the same line as before with the additional estimation of error (4) component and interaction with all other factors.

Example 11.16 With a view to formulate optimum spacing schedule for Rabi Crop of different duration, an experiment was conducted in strip-plot design at a certain research station during the year 1981.

The treatments were :

Spacing (4)	**Varieties (5)**	No. of replications = 3
S_1 = 10 Cm x 10 Cm	V_1 = PR 202	Plot size 3.3m x 2.4m
S_2 = 10 Cm x 5 Cm	V_2 = V_2M - 2	
S_3 = 5 Cm x 5 Cm	V_3 = CR - 652	
S_4 = 10 Cm. Solid rows.	V_4 = VR/Fa - 1	
	V_5 = AKP - 2.	

[illegible]nt and the yield in kg/plot is given below :

Rep—1

	S_2	S_1	S_4	S_3
[illegible]	4.20	1.80	3.32	2.94
[illegible]	3.75	3.29	1.38	3.29
[illegible]	1.14	3.48	3.10	4.24
[illegible]	3.75	4.82	4.67	4.14
[illegible]	3.60	3.34	3.95	1.54

Rep—2

	S_3	S_1	S_4	S_2
V_4	1.50	2.45	2.70	3.45
V_1	3.90	2.84	3.77	2.84
V_5	3.20	3.50	1.79	3.45
V_2	3.45	1.80	3.20	3.00
V_3	2.20	3.83	2.59	1.95

Rep—3

	S_4	S_1	S_2	S_3
V_2	3.05	4.25	2.59	1.49
V_4	3.30	2.84	2.70	3.50
V_5	1.89	3.29	3.27	3.30
V_3	3.45	1.09	3.29	3.05
V_1	2.84	2.40	1.18	2.50

Solution : Here, C. F. = 530.6211 and Total S. S. = 578.0275 - C. F. = 47.4069

Table-11.40
Variety x Replication Table

Variety	Rep—1	Rep—2	Rep—3	Total
V_1	17.38	13.35	8.92	39.65
V_2	12.26	11.45	11.38	35.09
V_3	11.97	10.57	10.88	33.42
V_4	12.48	10.10	12.34	34.87
V_5	11.71	11.94	11.75	35.40
Total	65.75	57.41	55.27	178.43

$$\text{Total S. S. (Variety x Rep. Table)} = \frac{17.38^2 + ... + 11.75^2}{4} - \text{C. F.}$$

$= 542.6722 - 530.6211 = 12.0511.$

$$\text{S. S. due to Variety} = \frac{39.65^2 + ... + 35.40^2}{12} - \text{C. F.} = 532.4503 - \text{C. F.} = 1.8\text{[illegible]}$$

S. S. due to Replication = $\frac{65.75^2 + ...+ 55.27^2}{20}$ - C. F.=533.6872 - C. F.=3.0661.

(Var x Rep) Interaction S. S. = 12.0511 - 1.8292 - 3.0661 = 7.1558 (E_1).

Table-11.41

Spacing x Replication Table

Spacing	Rep—1	Rep—2	Rep—3	Total
S_1	16.73	14.43	13.87	45.02
S_2	16.44	14.69	13.84	44.97
S_3	16.16	14.25	13.03	43.44
S_4	16.42	14.05	14.53	45.00
Total	65.75	57.41	55.27	178.43

Total S. S. (Spacing x Rep. Table) = $\frac{16.73^2 +......+ 14.53^2}{5}$ - C. F.

= 533.9900 - C. F. = 3.3690.

S. S. due to Spacing = $\frac{45.02^2 +...... + 45.00^2}{15}$ - C. F. = 530.7423 - C. F. = 0.1212.

(Spacing x Rep) Interaction S. S. = 3.3690 - 3.0661 - 0.1212 = 0.1817 = (E_2).

Table-11.42

Variety x Spacing Table

Variety	S_1	S_2	S_3	S_4	Total
V_1	10.06	9.09	9.22	11.28	39.65
V_2	7.85	8.69	8.98	9.57	35.09
V_3	8.40	6.14	9.74	9.14	33.42
V_4	8.63	10.55	5.74	9.95	34.87
V_5	10.08	10.50	9.76	5.06	35.40
Total	45.02	44.97	43.44	45.00	178.43

Total S. S. (Var. x Spa. Table) = $\frac{10.06^2 +.........+ 5.06^2}{3}$ - C. F.

(Var. x Spa.) Interaction S. S. = 16.8977 - 0.1212 - 1.8292 = 14.9473

Error S. S. (E_3) = 47.4069 - 1.8292 - 3.0661 - 7.1558 - 0.1212

- 0.1817 - 14.9473 = 20.1056

H_0 : The effect of all the spacing are equal.

Table-11.43

ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F	5% F
Replication	2	3.0661	1.5331		
Variety	4	1.8282	0.4573		
Rep. x Var. (E_1)	8	7.1558	0.8945		
Spacing	3	0.1212	0.0404	1.333	4.75
Rep. x Spa. (E_2)	6	0.1817	0.0303		
Var. x Spa.	12	14.9473	1.2456		
Error (E_3)	24	20.1056	0.8377		
Total	59	47.4059			

Since the calculated value of F corresponding to spacing with 3 and 6 d.f. is smaller than the tabulated value of F with same d.f. at 5% level of significance, the calculated value of F is insignificant and the hypothesis may be accepted.

11.8 Nested or Heirarchial Design

In multifactorial experiments there may be situation like that the number of levels of one factor are same to the other factor but the level may not be indentical to the other. Such an arrangement with two and more factors gives us nested or heirarchial design.

Let us consider an example that an industry purchases raw material from three different suppliers. The industry wishes to determine whether the genuinity of the raw material is the same from each supplier. There are four batches of raw material available from each suppllier and three observations are considered in each batch.

The physical condition of the design is given below :

Supplier	1				2				3			
Batches	1	2	3	4	1	2	3	4	1	2	3	4
	y_{111}	y_{121}	y_{131}	y_{141}	y_{211}	y_{221}	y_{231}	y_{24}	y_{311}	y_{321}	y_{331}	y_{341}
	y_{112}	y_{122}	y_{132}	y_{142}	y_{212}	y_{222}	y_{232}	y_{24}	y_{312}	y_{322}	y_{332}	y_{342}
	y_{113}	y_{123}	y_{133}	y_{143}	y_{213}	y_{223}	y_{233}	y_{24}	y_{313}	y_{323}	y_{333}	y_{343}

This is a two-stage nested design with batches nested within suppliers. It should be remembered that batch 1 or 2 etc. is not crossed with other factors i. e. batch 1 ofsuppliers 1 etc. is not same of batch 1 of supplier 2 and so on. Therefore the batches may be renumbered as 1, 2, 3, 4, for supplier 1 ; 5, 6, 7 and 8 for supplier 2 ; 9, 10, 11 and 12 for supplier 3, This is a balanced nested design, since there are an equal number of levels of one factor with in each level of the other factor and equal number of replicates. Since every level of one factor does not appear with every level of the other factors there can be no interaction between the two factors.

Analysis : Let y_{ijk} be the kth observation corresponding to the jth level of one factor B within ith level of the other factor A,

$i = 1, 2,........p\ ;\ j = 1, 2,......q$ and $k = 1, 2,.......r.$

We calculate $y_{i..} = \sum_j\sum_k y_{ijk}$; $y_{ij.} = \sum_k y_{ijk}$ and $y_{...} = \sum_i y_{i..} = \sum_i\sum_j\sum_k y_{ijk}$

$$\text{Total S.S.} = \sum_i\sum_j\sum_k y_{ijk}^2 - \frac{y_{...}^2}{pqr}$$

$$\text{S. S. due to A} = \sum_i \frac{y_{i..}^2}{qr} - \frac{y_{...}^2}{pqr}$$

$$\text{S. S. due to B within A} = \sum_i\sum_j \frac{y_{ij.}^2}{r} - \sum \frac{y_{i..}^2}{qr}$$

S. S. due to error can be obtained by usual subtraction and gives results

$$= \sum_i\sum_j\sum_k y_{ijk}^2 - \sum_i\sum_j y_{ij.}^2$$

The analysis of variance table for the two-stage nested design for testing the null hypothesis, H_0 : The effects of all the levels of first factor are same, is given in Table-11.44.

Table-11.44

ANOVA TABLE

Source of variation	d.f.	Sum of square
A	$p-1$	$\sum_{j} \frac{y_{i..}^2}{qr} - \frac{y_{...}^2}{pqr}$
B within A	$p(q-1)$	$\sum_{i}\sum_{j} \frac{y_{ij.}^2}{r} - \sum_{j} \frac{y_{i..}^2}{qr}$
Error	$pq(r-1)$	By subtractions $= \sum_{i}\sum_{j}\sum_{k} y_{ijk}^2 - \sum_{i}\sum_{j} \frac{y_{ij.}^2}{r}$
Total	$pqr-1$	$\sum_{i}\sum_{j}\sum_{k} y_{ijk}^2 - \frac{y_{...}^2}{pqr}$

The conclusion can be drawn as usual.

Example 11.17 A company which buys raw material in batches from three different suppliers. We wish to determine that all the suppliers provide material of same purity. Four batches of row material are selected at random from each supplier and the determination of purity is made on each batch. The data in a two-stage nested design are given below. Analyse the data.

Supplier	1				2				3			
Batches	1	2	3	4	1	2	3	4	1	2	3	4
	94	91	91	94	94	93	92	93	95	91	94	96
	92	90	93	97	91	97	93	96	97	93	92	95
	93	89	94	93	90	95	91	95	93	93	95	94

Solution : Correction Factor (C. F.) = $\frac{(3359)^2}{36}$ = 313413.36

Total S. S. = 313559 - 313413.36 = 145.64.

Batch totals within supplier

Supplier (A)	1				2				3			
Batch	1	2	3	4	1	2	3	4	1	2	3	4
Total	279	270	278	284	275	285	276	284	285	277	281	285
Total	1111				1120				1128			

S. S. due to A = $\frac{1111^2 + 1120^2 + 1128^2}{12}$ - C. F. = 313425.42 - C. F. = 12.06.

S. S. due to B (within A) = $\frac{279^2 + \dots\dots + 285^2}{3}$ - 313425.42

= 313501.00 - 313425.42 = 75.58

S. S. due to Error = 145.64 - 12.06 - 75.58 = 58

H_0 : All suppliers provide material of same purity.

Table-11.45

ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.	F	5% F
A	2	12.06	6.03	2.5	3.40
B(A)	9	75.58	8.40		
Error	24	58.00	2.42		
Total	35				

The calculated value of F with (2,24) d.f. is 2.5 which is smaller than the tabulated value of F with same d.f. at 5% level of significance. Hence the calculated value of F is insignificant and the hypothesis may be accepted.

12. INDEX NUMBER

12.1 Introduction

Index numbers are statistical devices designed to measure the relative change in the level of a phenomenon (variable or a group of variables) with respect to time, geographical location or other characteristics, such as income, production, expenditure, export, import, etc. In other words, index numbers are the numbers which indicate the value of a variable at any given date called the 'current period' as percentage of the value of that variable at some standard date called the 'base period'. The variable may be :

i) the prices of a particular set of commodities,

ii) the volume of trade, exports and imports, agricultural or industrial productions.

iii) the national income of a country or cost of living of persons belonging to a particular income group or profession.

12.2 Problem of Construction of Index Numbers

The construction of index number involves the following problems :

a) The purpose of index number.

b) Selection of commodities.

c) Selection of base.

d) Type of average to be used.

e) Selection of appropriate weight.

a) The Purpose of Index Number : If it is desired to construct an index of consumer's prices, we must know the class of consumers whose cost of living, we intend to measure and whether it is the cost of living of the middle class people, agriculturists or industrial workers. Such definiteness is necessary for the importance of various items consumed by the different categories of people may be very much different. It is always advisable as well as desirable to precisely know what we are going to measure as well as what purpose the measure is meant for.

b) **Selection of Commodities :** If the purpose of an index is to measure the cost of living of poor families we should select those commodities or items which are consumed by persons belonging to this group and due care should be taken not to include the goods which are not ordinarily consumed by the individuals of the selected families.

c) **Selction of Base :** The period with which the level of phenomena are made is termed as base period and the index for this period is always taken as 100. There are two types of base namely i) Fixed base and ii) Chain base.

i) **Fixed Base :** In fixed base method, the base period should be normal i. e. a period free from all sorts of abnormalities, such as economic depression, labour strikes, war, floods, earth-quake, etc.

The base period should not be too distant from the current period. Since index numbers are essential tools in business, planning and in formulation of executive decisions and hence the base period should not be too far back relative to current period. But the base period should be entirely different from the current period. Again the pattern of consumption of commodities may change appreciably if the base period is very far away from the current period.

ii) **Chain Base :** In the chain base method, the whole series of index number is not derived to any one base period, but the indices for different years are derived by relating each year's value to that of the immediately preceeding year, the indices so obtained are called link relative index numbers. Frequently, these link relatives are chain together to a common base. Such indices are known as chain indices. The chain base method provides for the inclusion of new items and deletion of old ones in order to make the index more representative.

d) **Type of Average to be Used :** Since index numbers are specialised averages, a judicious choice of average to be used in their construction is of great importance. Usually the averages namely i) Arithmetic mean, ii) Geometric mean and iii) Median are used.

Median, though easiest to calculate of all the three, completely ignores the extreme observations while arithmetic mean, though easy to calculate, is unduly affected by extreme observations. Moreover, neither arithmetic mean nor median are reversible. Geometric mean gives equal weights to equal ratios of change.

It does not give undue weightage to extreme observations. Geometric mean based indices are reversible. Hence geometric mean is the most appropriate average to be used.

e) **Selection of Appropriate Weights :** Genearally, for the construction of cost of living indices, various commodities such as wheat, rice, fuel, clothing etc. included in the index are not of equal importance, proper weights should therefore be attached to them to take into account for their relative importance.

12.3 Calculation of Index Numbers

Some simple but useful ways of calculating index numbers are given below :

A) **Simple Aggregate Method :** This method consists in expressing aggregate of prices in any year as a percentage of their aggregate in the base year. This price (or quantity) index for the ith year (i = 1, 2,......., n) as compared to the base year (i = 0) is given by

$$P_{oi} = \frac{\sum_{j=1}^{r} P_{ij}}{\sum_{j=1}^{r} P_{oj}} \times 100 \qquad \text{.............(12.1)}$$

where, P_{oi} = Price index of the ith (i = 1, 2,......., n) year with respect to base year,

P_{ij} = Price of the ith year of the jth (j = 1, 2,......., r) commodity,

and P_{oj} = Price of the base year of the jth commodity.

$$\text{And, } Q_{oi} = \frac{\sum_{j=1}^{r} q_{ij}}{\sum_{j=1}^{r} q_{oj}} \times 100 \qquad \text{.............(12.2)}$$

where, Q_{oi} = Quantity index of the ith year with respect to the base year,

q_{ij} = Quantity of the jth commodity in the ith year,

q_{oj} = Quantity of the jth commodity in the base year.

Defects of this method are :

i) The prices of the various commodities may be in different units, e. g. per litre, per metre, per quintal etc.

ii) The relative importance of various commodities are neglected.

Example 12.1 Construct index number of prices of 1990 taking 1985 as the base from the following data using simple aggregate method.

Commodity	Price in 1985 in Taka	Price in 1990 in Taka
Rice	10.5 per kg.	15.5 per kg.
Wheat	5.5 per kg.	6.5 per kg.
Cloth	5.5 per metre	7.0 per metre
Sugar	20.5 per kg.	27.5 per kg.
Milk	8.0 per kg.	14.5 per kg.

Solution :

Commodity	Price in 1985 in Taka. P_0	Price in 1990 in Taka P_1
Rice	10.5	15.5
Wheat	5.5	6.5
Cloth	5.5	7.0
Sugar	20.5	27.5
Milk	8.0	14.5
Total	50.0	71.0

Therefore, price index number of 1990 using 1985 as base is

$$P_{oi} = \frac{\Sigma P_{ij}}{\Sigma P_{oj}} \times 100 = \frac{71}{50} \times 100 = 142.0$$

B) Weighted Aggregate Method : This method provides for the different commodities to exert their influence in the index number by assigning appropriate weights to each. Usually the quantity consumed, sold or marketed in the base year are used as weights. If w_j is the weight associated with the jth commodity then the weighted aggregate price index is given by,

$$P_{oi} = \frac{\sum_{j=1}^{r} w_j P_{ij}}{\sum_{j=1}^{r} w_j P_{oj}} \times 100 \qquad \text{.............(12.3)}$$

where, P_{ij} and P_{oj} are as expressed in (12.1)

By the use of different types of weights, a number of formulae have emerged for the construction of index number.

12.3.1 Laspeyre's Price Index : If we take $w_j = q_{oj}$ in (12.3) i.e. if the base year quantities used as weights, the method is called Laspeyre's method and the formula is,

$$P_{oi}\,(La) = \frac{\sum_{j=1}^{r} P_{ij}\, q_{oj}}{\sum_{j=1}^{r} P_{oj}\, q_{oj}} \times 100 \qquad \text{.................(12.4)}$$

where, the notations are expressed earlier.

12.3.2 Paasche's Price Index : By taking current year quantities as weights, i.e. $w_j = q_{ij}$ in (12.3) the method is known as Paasche's method and the formula is,

$$P_{oi}\,(Pa) = \frac{\sum_{j=1}^{r} P_{ij}\, q_{ij}}{\sum_{j=1}^{r} P_{oj}\, q_{ij}} \times 100 \qquad \text{...........(12.5)}$$

12.3.3 Drobish -Bowley Price Index : This method is the arithmetic mean of the Laspeyre's and Paasche's price indices and is given by,

$$P_{oi}\,(DB) = \frac{1}{2}\left[\frac{\Sigma P_{ij} q_{oj}}{\Sigma P_{oj} q_{oj}} + \frac{\Sigma P_{ij} q_{ij}}{\Sigma P_{oj} q_{ij}}\right] \times 100 \qquad \text{............(12.6)}$$

12.3.4 Marshall-Edgeworth Price Index : If $w_j = \frac{1}{2}(q_{oj} + q_{ij})$ in (12.3) i.e. if weights are the arithmetic mean of the base year quantities and the current year quantities, the method is known as Marshall-Edgeworth method and the formula is given by,

$$P_{oi}(ME) = \frac{\Sigma P_{ij}\left(\frac{q_{oj} + q_{ij}}{2}\right)}{\Sigma P_{oj}\left(\frac{q_{oj} + q_{ij}}{2}\right)} \times 100$$

$$\text{or, } P_{oi}\,(ME) = \frac{\Sigma P_{ij}\,(q_{oj} + q_{ij})}{\Sigma P_{oj}\,(q_{oj} + q_{ij})} \times 100 \qquad \text{..............(12.7)}$$

12.3.5 Walsch Price Index : If the weights are the geometric mean of the base year quantities and the current year quantities, the method is known as Walsch method and the formula is given by

$$P_{oi}(Wa) = \frac{\sum P_{ij}\sqrt{q_{oj}q_{ij}}}{\sum P_{oj}\sqrt{q_{oj}q_{ij}}} \times 100 \qquad \text{..........(12.8)}$$

12.3.6 Irving Fisher's Ideal Price Index : The geometric mean of Laspeyre's and Paasche's formula is known as Fisher's ideal price index and is given,

$$P_{oi}(F) = [P_{oi}(L_a) \times P_{oi}(P_a)]^{\frac{1}{2}}$$

$$= \left[\frac{\sum P_{ij}q_{oj}}{\sum P_{oj}q_{oj}} \cdot \frac{\sum P_{ij}q_{ij}}{\sum P_{oj}q_{ij}}\right]^{\frac{1}{2}} \times 100 \qquad \text{............(12.9)}$$

12.3.7 Kelly's Price Index : If in (12.3) the weights w_j are not the quantities which refer to some period (not necessarily the base year or current year) and are kept constant for all periods, the method is known as Kelly's method.

Note : 1) Quantity Index Number : In the above formula (12.4) to (12.9) we concentrated ourselves on price index numbers. By interchanging the prices (P_{ij}) and quiantities (q_{ij}) in the above formulae, we get corresponding formulae for the calculation of quantity index numbers, which reflect the change in the volume of quantity or production.

2) Value Index Number : Value index numbers are given by the aggregate expenditure for any given year expressed as a percentage of the same in the base year. Thus $V_{oi} = \frac{\sum P_{ij}q_{ij}}{\sum P_{oj}q_{oj}} \times 100$ ....(12.10)

where V_{oi} is the value index and the other notations are as usual.

Example 12.2 Construct index number of prices from the following data by using : a) Laspeyre's method b) Paasche's method c) Marshall-Edgeworth's method d) Fisher's method

	1985		1990	
Commodity	Price (Taka)	Quantity	Price (Taka)	Quantity
Rice	10.5	3	15.5	4
Wheat	5.5	2	6.5	3
Cloth	5.5	5	7.0	7
Sugar	20.5	1	27.5	2
Milk	8.0	1	14.5	2

Solution :

	1985		1990					
Comodity	Price P_o	Quan. q_o	Price P_1	Quan. q_1	P_oq_o	P_oq_1	P_1q_o	P_1q_1
Rice	10.5	3	15.5	4	31.5	42.0	46.5	62.0
Wheat	5.5	2	6.5	3	11.0	16.5	13.0	19.5
Cloth	5.5	5	7.0	7	27.5	38.5	35.0	49.0
Sugar	20.5	1	27.5	2	20.5	41.0	27.5	55.0
Milk	8.0	1	14.5	2	8.0	16.0	14.5	29.0
Total					98.5	154.0	136.5	214.5

Index number for 1990 with base 1985 by using :

(a) Laspeyre's method, P_{oi} $\dfrac{\sum_{j=1}^{5} P_{ij}q_{oj}}{\sum_{j=1}^{5} P_{oj}q_{oj}} \times 100 = \dfrac{136.5}{98.5} \times 100 = 138.58$

(b) Paasche's method, $P_{oi} = \dfrac{\sum_{j=1}^{5} P_{ij}\, q_{ij}}{\sum P_{oj}\, q_{ij}} \times 100 = \dfrac{214.5}{154} \times 100 = 139.29$

(c) Marshall-Edgeworth's method, $P_{oi} = \dfrac{\sum_{j=1}^{5} P_{ij}\,(q_{oj} + q_{ij})}{\sum_{j=1}^{5} P_{oj}\,(q_{oj} + q_{ij})} \times 100$

$$= \frac{\sum_{j=1}^{5} P_{ij}\, q_{oj} + \sum_{j=1}^{5} P_{ij}\, q_{ij}}{\sum_{j=1}^{5} P_{oj}\, q_{oj} + \sum_{j=1}^{5} P_{oj}\, q_{ij}} \times 100 = \frac{136.5 + 214.5}{98.5 + 154} \times 100$$

$$= \frac{351}{252.5} \times 100 = 139.01$$

(d) Fisher's method, $$P_{oi} = \sqrt{\frac{\sum\limits_{j=1}^{5} P_{ij}q_{oj}}{\sum\limits_{j=1}^{5} P_{oj}q_{oj}} \times \frac{\sum\limits_{j=1}^{5} P_{ij}q_{ij}}{\sum\limits_{j=1}^{5} P_{oj}q_{ij}}} \times 100$$

$$= \sqrt{\frac{136.5}{98.5} \cdot \frac{214.5}{154}} \times 100 = 138.93$$

12.4 Simple Average of Price Relative Method

As the name implies, this method consists of finding price relatives and averaging them expressed in percentage. A price relative is the ratio of price of the commodity in the current year divided by the price of the same commodity in the base year. Symbolically price relative is $\frac{P_{ij}}{P_{oj}}$.

The next step is to average this price relatives of each current year and then express in percentage to obtain the index number.

For the purpose of the averages any one measure of central location, such as mean, median, geometric mean may be used. Therefore, the simple average of price relative index number is

$$P_{oi}\,(A.\,M) = \frac{\sum\limits_{j=1}^{r} \frac{P_{ij}}{P_{oj}}}{N} \times 100 \qquad \text{.............(12.11)}$$

When arithmetic mean is taken, N is the number of commodities and

$$P_{oi}\,(G.\,M.) = \left(\prod_{j=1}^{r} \frac{P_{ij}}{P_{oj}}\right)^{\frac{1}{N}} \times 100 \qquad \text{.............(12.12)}$$

When geometric mean is taken, N is the number of commodities.

12.4.1 Weighted Average of Price Relatives : For the obvious short coming of the simple average of relatives is that each relative irrespective of the importance of the commodity it presents, influence the index number for a given year. If w_j is the weight given to jth commodity, then the general formulae of index numbers obtained on taking the weighted average of price relatives become :

$$P_{oi}\,(A.\,M.) = \frac{\sum\limits_{j=1}^{r} wj\left(\frac{P_{ij}}{P_{oj}}\right)}{\sum\limits_{j=1}^{r} w_j} \times 100 \qquad \text{........(12.13)}$$

$$P_{oi}\,(G.\,M.) = \left[\prod_{j=1}^{r}\left(\frac{P_{ij}}{P_{oj}}\right)^{w_j}\right]^{\frac{1}{\Sigma w_j}} \times 100 \qquad \text{............(12.14)}$$

If the base year values are taken as weights, i.e., $w_j = P_{oj}q_{oj}$,

we get from (12.13)

$$P_{oi}\,(A.\,M.) = \frac{\Sigma P_{ij}q_{oj}}{\Sigma P_{oj}\,q_{oj}} \times 100 \qquad \text{............(12.15)}$$

which is nothing but Laspeyre's formula as obtained in (12.4)

If we take the values obtained by multiplying the current year quantities and the base year prices as weight i.e. we take $w_j = P_{oj}q_{ij}$, we get from (12.13)

$$P_{oi}\,(A.M.) = \frac{\Sigma P_{ij}q_{ij}}{\Sigma P_{oj}q_{ij}} \times 100$$

which is Paasche's formula as obtained in (12.5).

Example 12.3 The price of four different commodities for 1986 and 1990 are given below. Calculate the index number for 1990 with 1986 as base using (i) the simple average of price relative method (ii) the weight average of price relative method.

Commodity	Weight	Prices in Taka			
		1986		1990	
Rice	3	11.0	per kg.	15.5	per kg.
Wheat	3	5.0	per kg.	6.5	per kg.
Cloth	4	5.5	per metre	7.0	per metre
Sugar	1	22.5	per kg.	27.5	per kg.

Solution :

Commodity	Weight w	Base year Price (1986) P_o	Current year Price (1990) P_1	Price Relative $\frac{P_1}{Po}$	$w.\frac{P_1}{Po}$
Rice	3	11.0	15.5	1.409	4.227
Wheat	3	5.0	6.5	1.300	3.900
Cloth	4	5.5	7.0	1.273	5.092
Sugar	1	22.5	27.5	1.222	1.222
	11			5.204	14.441

i) Simple average of price relative index is given by,

$$P'_{o1} = \frac{\Sigma \frac{P_{1j}}{P_{oj}}}{N} \times 100 = \frac{5.204}{4} \times 100 = 130.1$$

ii) Weighted average of price relative index is given by,

$$P_{o1} = \frac{\sum_{j=1}^{4} w_j \frac{P_{1j}}{P_{oj}}}{\sum_{j=1}^{4} w_j} = \frac{14.441}{11} \times 100 = 131.28$$

12.5 Tests of Index Numbers.

The following are the tests commonly used for the test of index numbers.

A) Time Reversal Test.

B) Factor Reversal Test.

C) Circular Test.

A) Time Reversal Test : The test is that the index numbers of current year to the base year should be the reciprocal of the index number of base year to the current year. Symbolically,

$$P_{oi} = \frac{1}{P_{io}}$$

or, $P_{oi} \cdot P_{io} = 1$

For example, if we take the Laspeyre's formula

$$P_{oi}(L_a) = \frac{\Sigma P_{ij} q_{oj}}{\Sigma P_{oj} q_{oj}} \text{ Also we get, } P_{io}(L_a) = \frac{\Sigma P_{oj} q_{ij}}{\Sigma P_{ij} q_{ij}}$$

$$\therefore\ P_{oi}.(L_a)\ P_{io}\,(L_a) = \frac{\Sigma P_{ij} q_{oj}}{\Sigma P_{oj} q_{oj}} \cdot \frac{\Sigma P_{oj} q_{ij}}{\Sigma P_{ij} q_{ij}} \neq 1$$

Hence Laspeyre's formula does not satisfy time reversal test. Similarly it can be shown that Paasche's formula does not satisfy this test. For Fisher's Ideal Formula,

$$P_{oi}\,(F) = \left[\frac{\Sigma P_{ij} q_{oj}}{\Sigma P_{oj} q_{oj}} \cdot \frac{\Sigma P_{ij} q_{ij}}{\Sigma P_{oj} q_{ij}}\right]^{\frac{1}{2}} \quad \text{and } P_{io}\,(F) = \left[\frac{\Sigma P_{oj} q_{ij}}{\Sigma P_{ij} q_{ij}} \cdot \frac{\Sigma P_{oj} q_{oj}}{\Sigma P_{ij} q_{oj}}\right]^{\frac{1}{2}}$$

$$\therefore\ P_{oi}\,(F).\,P_{io}\,(F) = 1.$$

Hence Fisher's ideal index satifies time reversal test. It can be easily shown that simple aggregate index and Marshall-Edgeworth index (with out the factor 100) also satisfy this test.

B) Factor Reversal Test : The factor reversal test requires that the product of a price index and the corresponding quantity index should be equal to value index, the indices being expressed in ratio. Symbolically

$$P_{oi} \cdot Q_{oi} = \frac{\Sigma v_{ij}}{\Sigma v_{oj}} = \frac{\Sigma P_{ij} q_{ij}}{\Sigma P_{oj} q_{oj}}.$$

For example,

$$P_{oi}\,(F) = \left[\frac{\Sigma P_{ij} q_{oj}}{\Sigma P_{oj} q_{oj}} \cdot \frac{\Sigma P_{ij} q_{ij}}{\Sigma P_{oj} q_{ij}}\right]^{\frac{1}{2}} \quad \text{and } Q_{oi}\,(F) = \left[\frac{\Sigma q_{ij} P_{oj}}{\Sigma q_{oj} P_{oj}} \cdot \frac{\Sigma q_{ij} P_{ij}}{\Sigma q_{oj} P_{ij}}\right]^{\frac{1}{2}}$$

$$\therefore\ P_{oi}.(F)\ Q_{oi}\,(F) = \frac{\Sigma P_{ij} q_{ij}}{\Sigma P_{oj} q_{oj}} \text{ (on simplification)}$$

Hence Fisher's ideal index satisfies factor reversal test and none of other formulae satisfies the factor reversal test.

Remarks :

(1) In varification of these tests various formulae are taken without the factor 100.

(2) Since Fisher's index satisfies both time reversal test and factor reversal tests, it is termed as ideal index number.

C) Circular Test : This test is based on the shift-ability of the base and is an extention of the time reversal test. The test is that

$P_{oi} \cdot P_{ij} \cdot P_{jo} = 1,\ i \neq j \neq 0.$

or, $P_{ab} \cdot P_{bc} \cdot P_{ca} = 1,\ a \neq b \neq c.$

This test is satisfied only by the indices based on

i) Simple geometric mean of price relative.

ii) Kelly's fixed weight method.

12.6 Cost of Living Index or Consumer's Price Index

Cost of living index numbers are constructed to study the effects of changes in the prices of a basket of goods and services on the purchasing power of a particular class of people during current period as compared with some base period. Change in the cost of living of an individual between two periods means the change in his money income which will be necessary for him to maintain the same standard of living in both periods. The consumption habits of people differ widely from class to class and even within the same class from region to region, the changes in the level of prices affect different classes differently and consequently the general price index number usually fail to reflect the effects of changes in the general prices level on the cost of living of different classes of people. Cost of living index numbers are therefore, compiled to get a measure of the general price movement of the commodities consumed by different classes of people.

For change in the cost of living may also arise from reasons other than price change and the cost of living does not measure such kind of change. From this point of view the cost of living index number should be called "Consumer's price index number."

12.6.1 Construction of Cost of Living Index Number : Cost of living number is constructed by the following formulae

a) Aggregate Expenditure Method or Weighted Aggregate Method.

b) Family Budget Method or Method of Weighted Relatives.

a) Aggregate Expenditure Method : In this method weights to be assigned to various commodities are provided by the quantities consumed in the base year. Thus in the usual notation cost of living index is given by,

$$P_{oi} = \frac{\Sigma P_{ij} q_{oj}}{\Sigma P_{oj} q_{oj}} \times 100$$

Note : This is nothing but Laspeyre's index.

b) Family Budget Method : In this method cost of living index is given by weighted average of price relatives, the weight being the values of quantities consumed in the base year. Thus in the usual notation cost of living index is given by ;

$$P_{oi} = \frac{\Sigma w_j \dfrac{P_{ij}}{P_{oi}}}{\Sigma w_j} \times 100, \text{ where } w_j = P_{oj} q_{oj}.$$

It is to be noted that cost of living index numbers by both the methods agree,

since $$\frac{\Sigma w_j \frac{P_{ij}}{P_{oj}}}{\Sigma w_j} \times 100 = \frac{\Sigma P_{oj} q_{oj} \frac{P_{ij}}{P_{oj}}}{\Sigma P_{oj} q_{oj}} \times 100 = \frac{\Sigma P_{ij} q_{oj}}{\Sigma P_{oj} q_{oj}} \times 100$$

Example 12.4 Construct the cost of living index for the year 1988 (Base 1984 = 100)

Commodity		Price in Taka		
	Unit	1984	1988	Weight
Rice	kg.	9.00	10.50	35%
Wheat	kg.	5.50	6.00	25%
Vegetables	kg.	2.50	3.50	20%
Meat	kg.	45.00	60.00	10%
Eggs	Dozon	5.50	7.50	10%

Solution : We prepare the folllowing table for calcu;lating cost of living index.

Commodity	Price in Taka 1984 P_o	1988 P_1	Price Relative P_1/P_o	Weight w	$w\frac{P_1}{P_o}$
Rice	9.00	10.50	1.667	35	58.345
Wheat	5.50	6.00	1.091	25	27.275
Vegetables	2.50	3.50	1.400	20	28.000
Meat	45.00	60.00	1.333	10	13.330
Eggs	5.00	7.50	1.364	10	13.640
Total				100	140.590

Cost of Living Index, $P_{o1} = \frac{\Sigma w_j \frac{P_{1j}}{P_{oj}}}{\Sigma w_j} = \frac{140.590}{100} \times 100 = 140.59$

Therefore, cost of living index for the year 1988 is 140.59 considering the base year 1984 = 100.